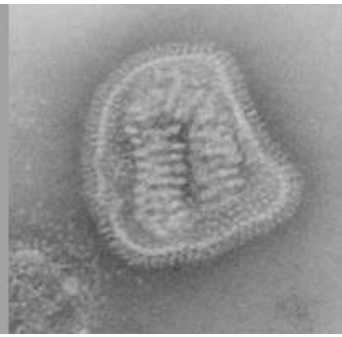
Prokaryotic and viral genomes. Eukaryotic organelles



Máté Varga
(mvarga@ttk.elte.hu)

Genome sizes

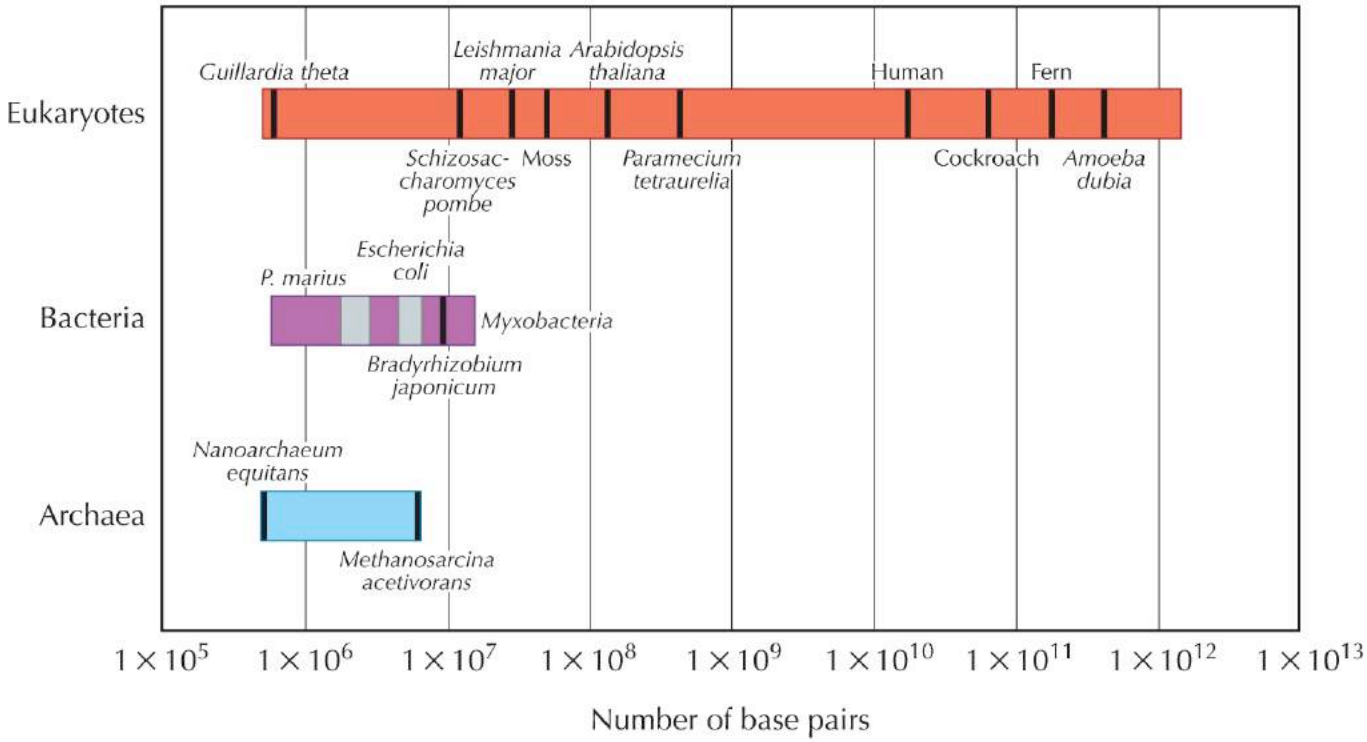
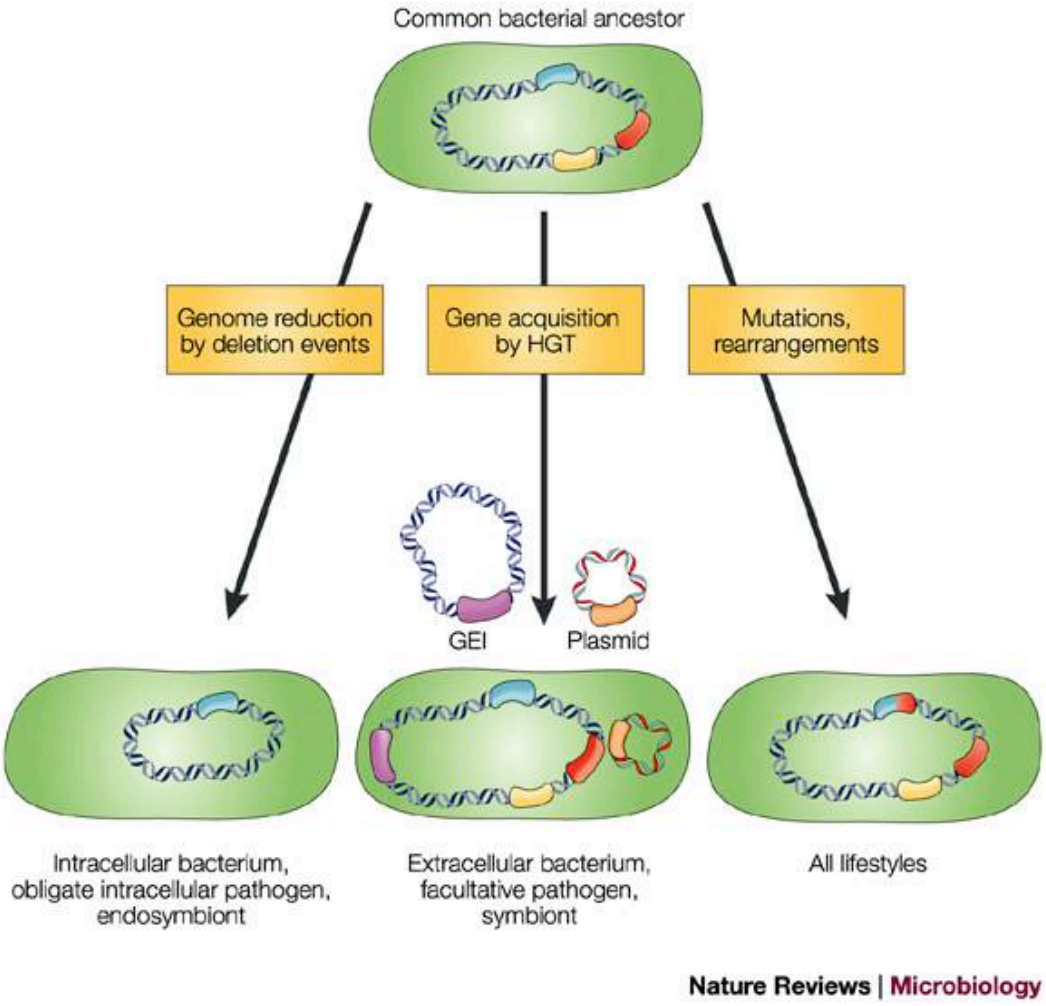


FIGURE 7.1. Genome sizes in the three domains of life. A selection of genome sizes and size ranges from specific groups of organisms is indicated.

7.1, adapted from Bentley S.D. et al., *Annu. Rev. Genet.* **38:** 771–791, © 2004 Annual Reviews, www.annualreviews.org, based on data from DOGS http://www.cbs.dtu.dk/databases/DOGS/

The size and the structure of the genome reflects the “lifestyle” of the bacteria

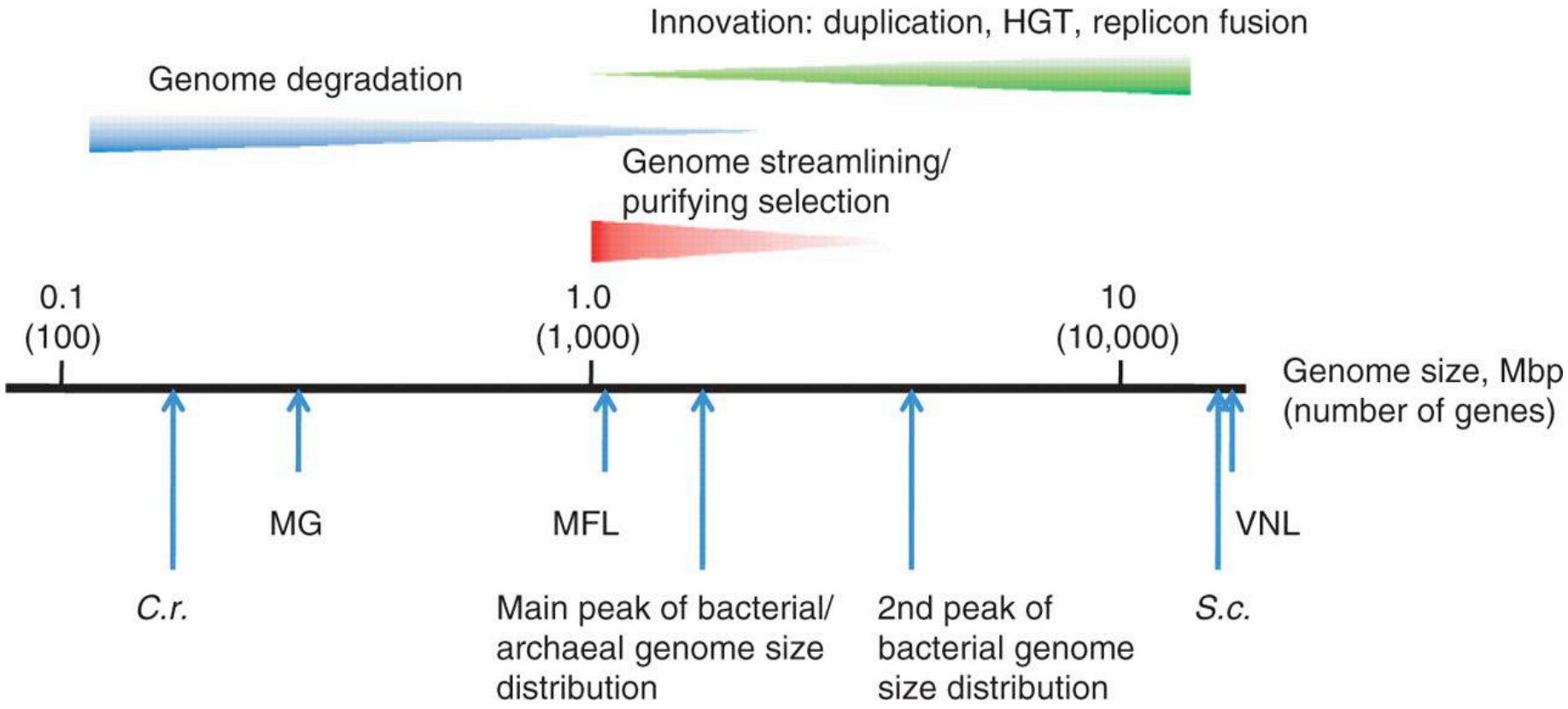


-a **stable environment** usually leads to genome reduction (genes not important for the given niche will be lost over time)

- a **changing environment** leads to larger genomes, as bacteria have to be ready for multiple possibilities (different circumstances need different set of genes for adaptation)

(Dobrindt et al. (2004) *Nat Rev Microb*)

Effects regulating the size of the bacterial genome



(Koonin and Wolf (2008) *Nuc Ac Res*)

Stable environment results in genome reduction

Table 1 General features of the genomes of *B. pertussis*, *B. parapertussis* and *B. bronchiseptica*

	B. pertussis	*B. parapertussis*	*B. bronchiseptica*
Size (bp)	4,086,186	4,773,551	5,338,400
G+C content (%)	67.72	68.10	68.07
Coding sequences	3,816	4,404	5,007
Pseudogenes	358 (9.4%)	220 (5.0%)	18 (0.4%)
Coding density (intact genes)	82.9%	86.6%	91.4%
Coding density (all genes)	91.6%	92.2%	92.0%
Average gene size (bp)	978	987	978
rRNA operons	3	3	3
tRNA	51	53	55
IS*481*	238	0	0
IS*1001*	0	22	0
IS*1002*	6	90	0
IS*1663*	17	0	0

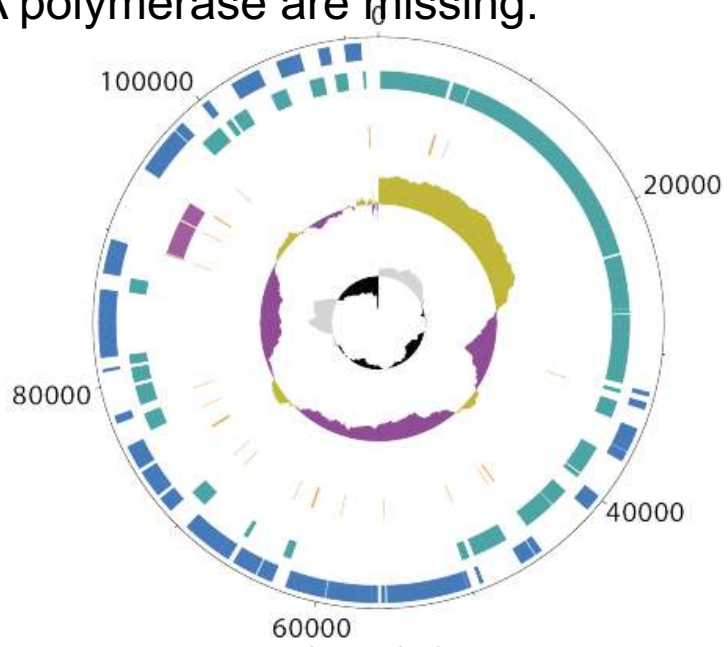
- the obligate human pathogen *B. pertussis* has lost 20% of its chromosome (*B. bronchiospetica* can infect multiple species)
- 10% of the remaining genes are pseudogenes
- a major force for gene inactivation is the expansion of an IS element

(Parkhill et al. (2003) *Nat Gen*)

The smallest (known) bacterial genome: *Nasuia deltocephalinicola*

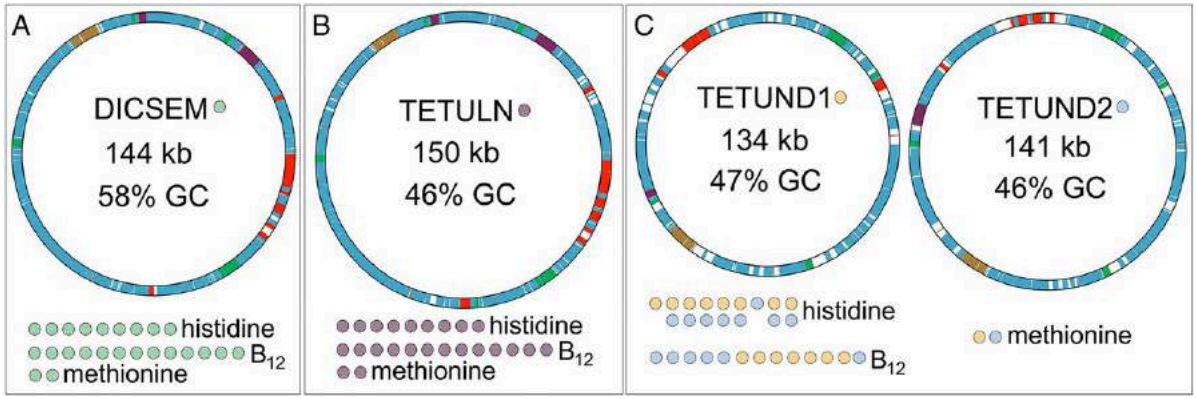
- Hemiptera feed on plant fluids, therefore they are missing 10 of the essential AAs from their diet - these are produced by obligate symbionts
- *Sulcia* sp. are the most important symbionts, but they produce only 8 of these missing AAs in some cicadas
-*Sulcia* have a reduced genome themselves – e.g. the symbiont of *Macrosteles quadrilineatus* has a 190 kb genome, with 190 protein coding genes, without DNA repair, oxidative phosphorylation, and even some components of the DNA polymerase are missing.

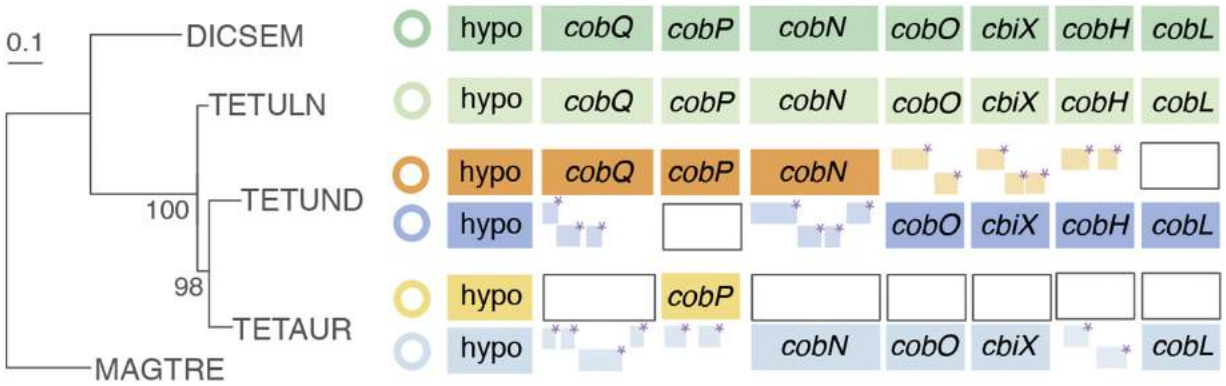
- Another symbiont of *M. quadrilineatus*, *Nasuia deltocephalinicola* has the smallest known genome
- This produce the two AAs not made by *Sulcia*: Met, His – there is barely any other metabolic pathway intact
- 112 Kb genome, 137 protein coding genes, alternative genetic code(UAS: STOP -> Trp)

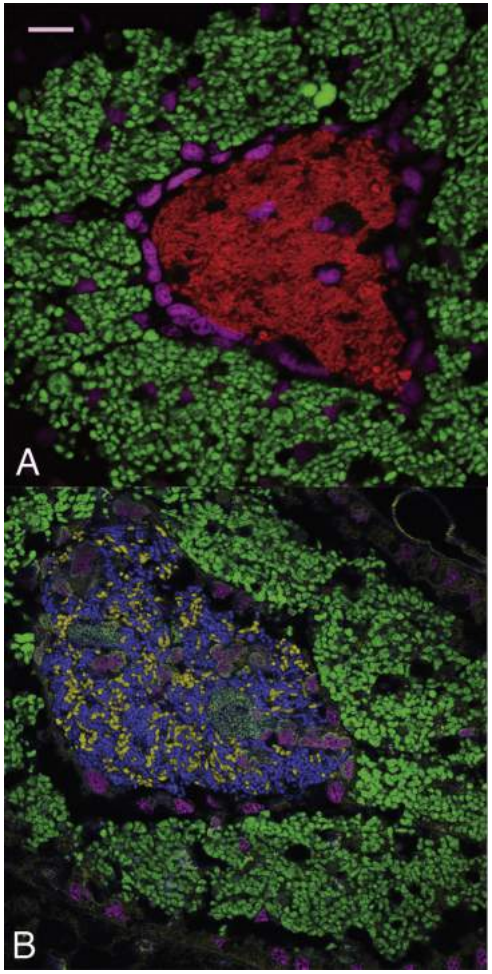


(Bennett and Moran (2013) *Genome Biol Evol)*

Complementary genome reduction in symbionts



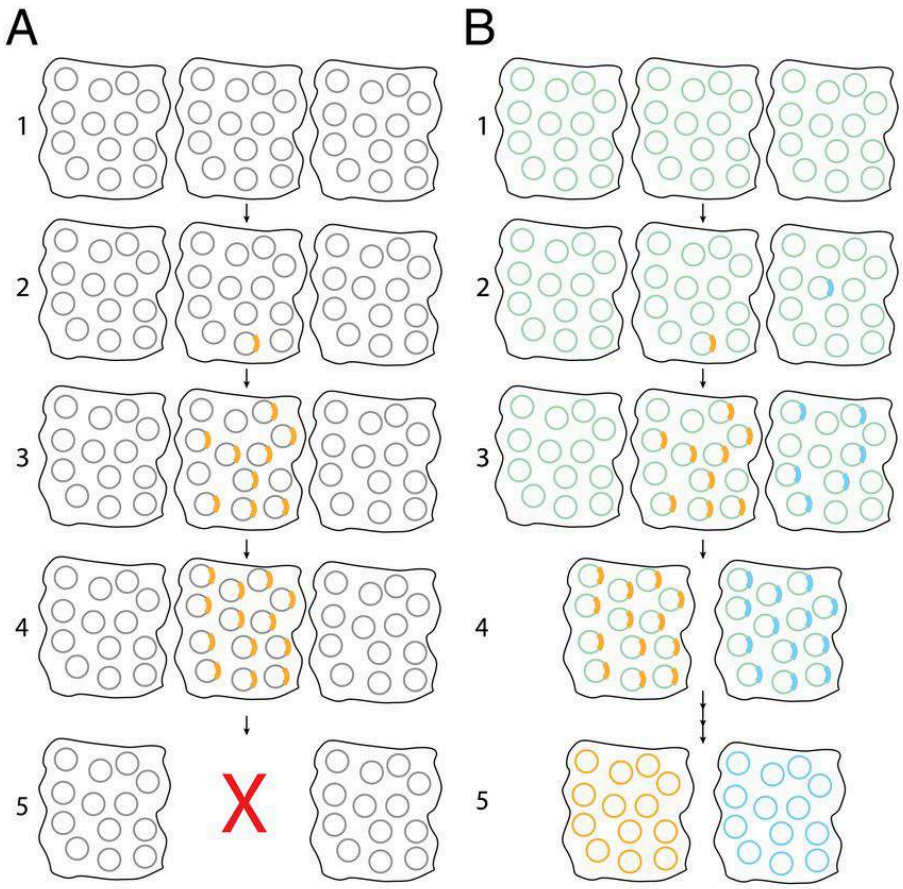




- *Hodgkinia* genomes from the cicada *Tettigades undata* can be assembled only on two chromosomes
- These genomes have complementary AA-synthesis pathways
- Only genome-based staining can distinguish them, their rRNA is almost identical

(Van Leuven et al. 2014 *Cell*, Campbell et al. 2014 *PNAS*)

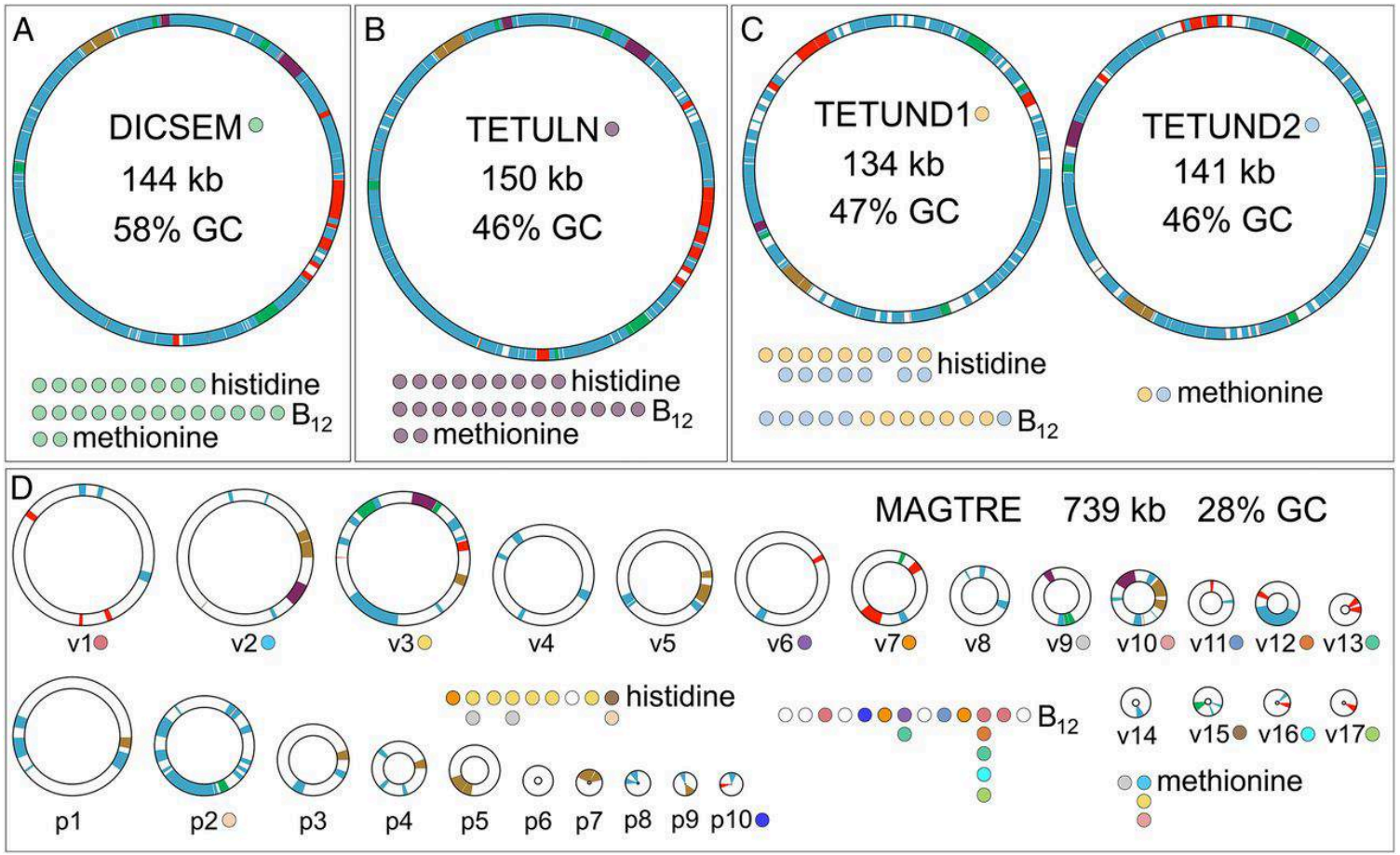
Complementary genome reduction in symbionts



- A: in some symbionts (pl. *Sulcia*) inactivating mutations appear (2) and become widespread through genetic drift (4), but finally natural selection discards them
- B: in the case of *Hodgkinia complementary* inactivating mutations appear, which can spread through the population, but if the original genome gets lost, these reduced genome symbionts will be dependent on each other

(Van Leuven et al. 2014 *Cell*, Campbell et al. 2014 *PNAS*)

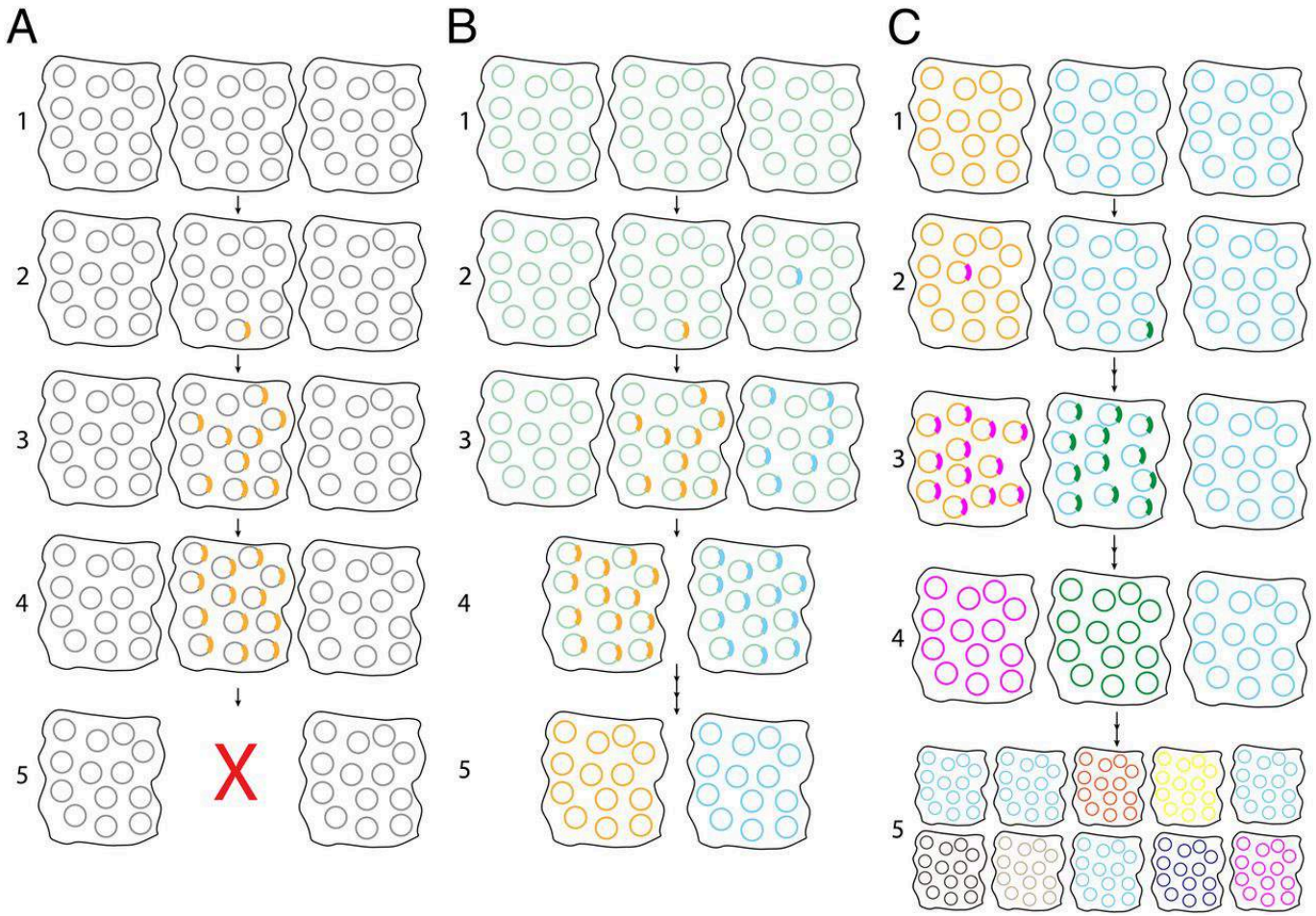
Complementary genome reduction in symbionts



- In *Magicicada tredicim* (MAGTRE) more than a dozen (min. 17) circular, complementary *Hodgkinia* “scaffolds” can be assembled. Some of these are almost certainly in different *Hodgkinia* cells.

(Campbell et al. 2014 *PNAS*)

Complementary genome reduction in symbionts

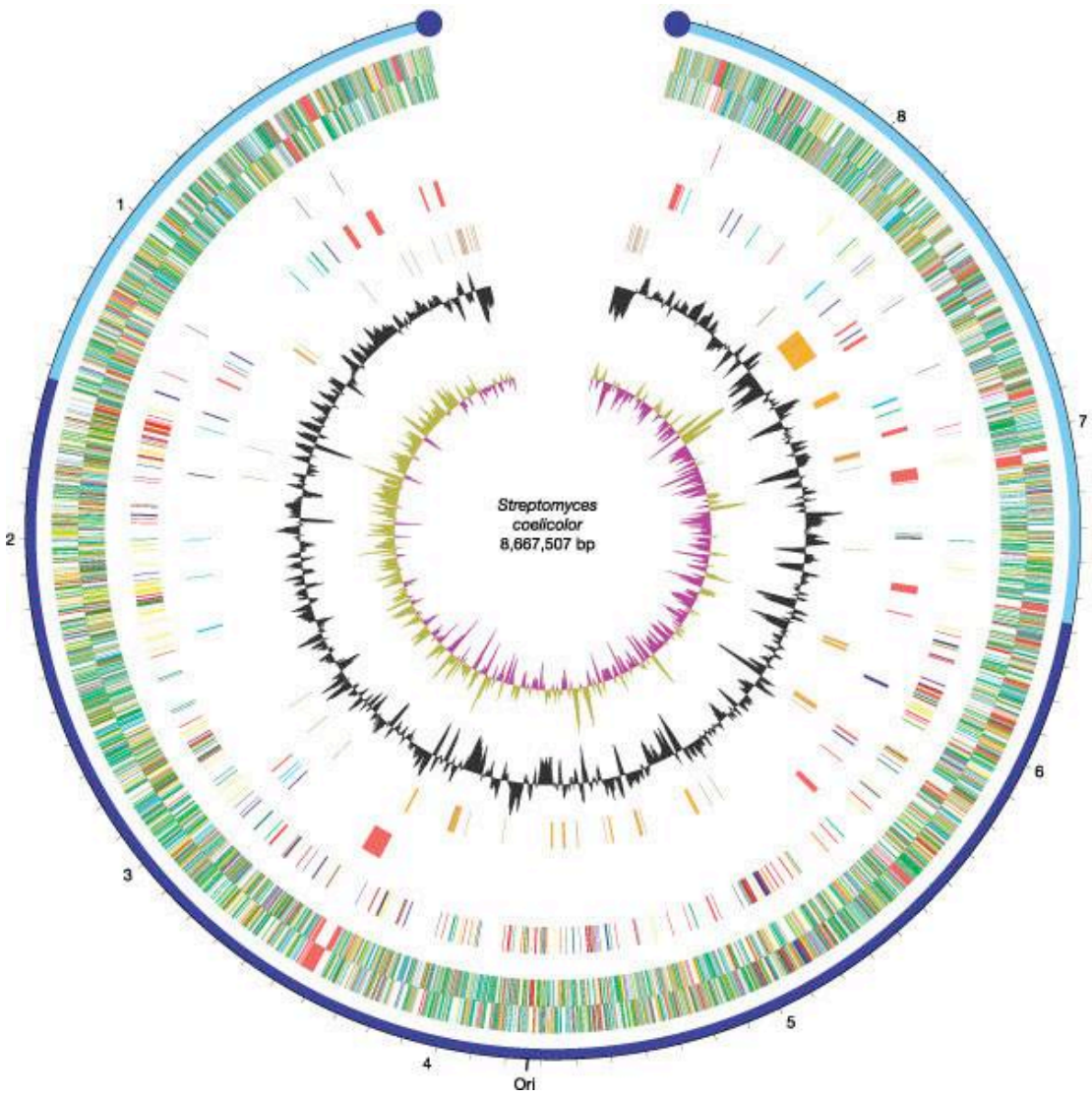


- In *Magicicada tredicim* the new genotypes segregated into further genotypes. Because of the complicated dependency network (the cicada and all the *Hodgkinia* lines are dependent on each other for survival), if a single *Hodgkinia* line gets lost, the whole ecosystem colapses.

(Campbell et al. 2014 *PNAS*)

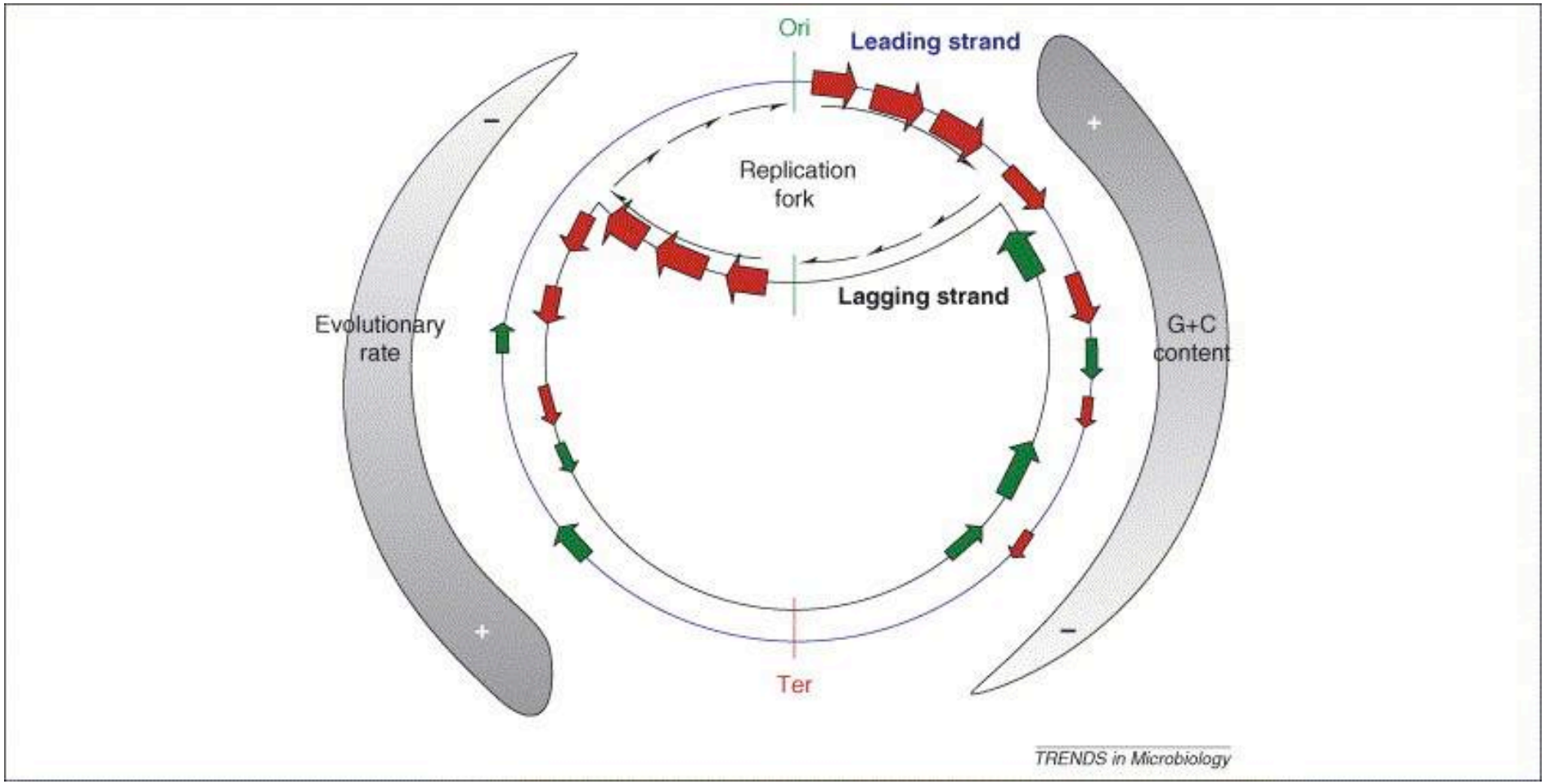
A changing environment requires a larger genome

- The genome of the soil species *Streptomyces coelicolor* is 8.7 Mb and codes for ~7800 proteins
- the essential genes are in the middle of the linear chromosome
- the non-essential “standby” genes are on the arms



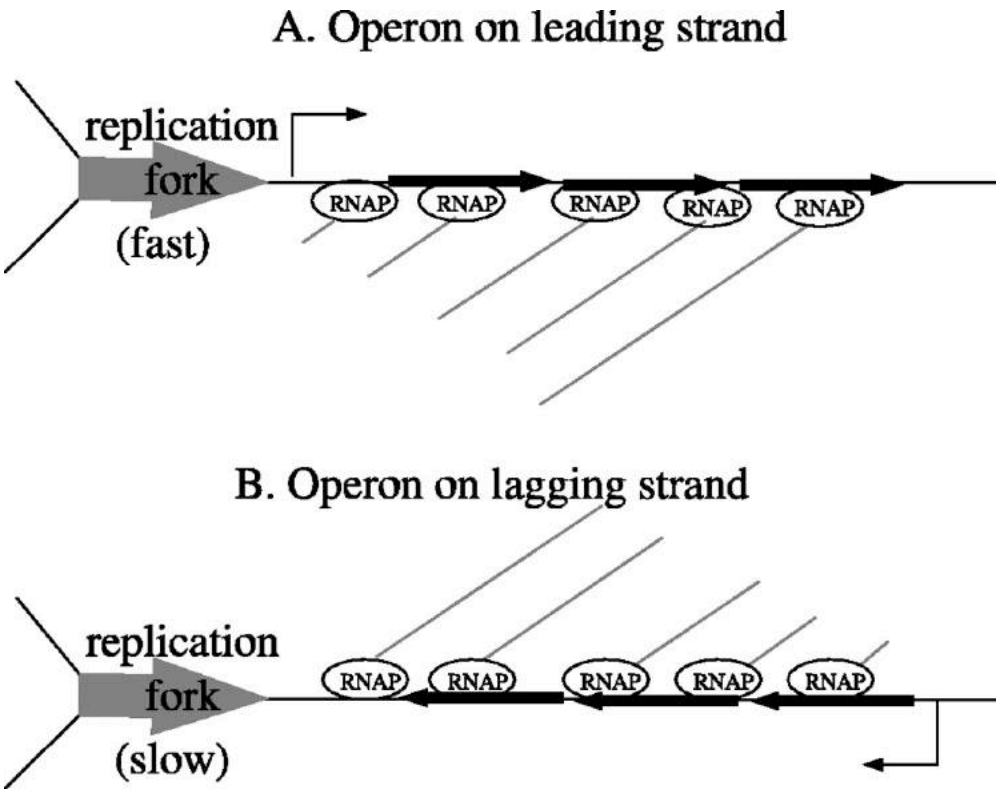
(Bentley et al. (2002) *Nature*)

Effects regulating the evolution of the bacterial chromosome



(Abby and Daubin (2007) *TiM*)

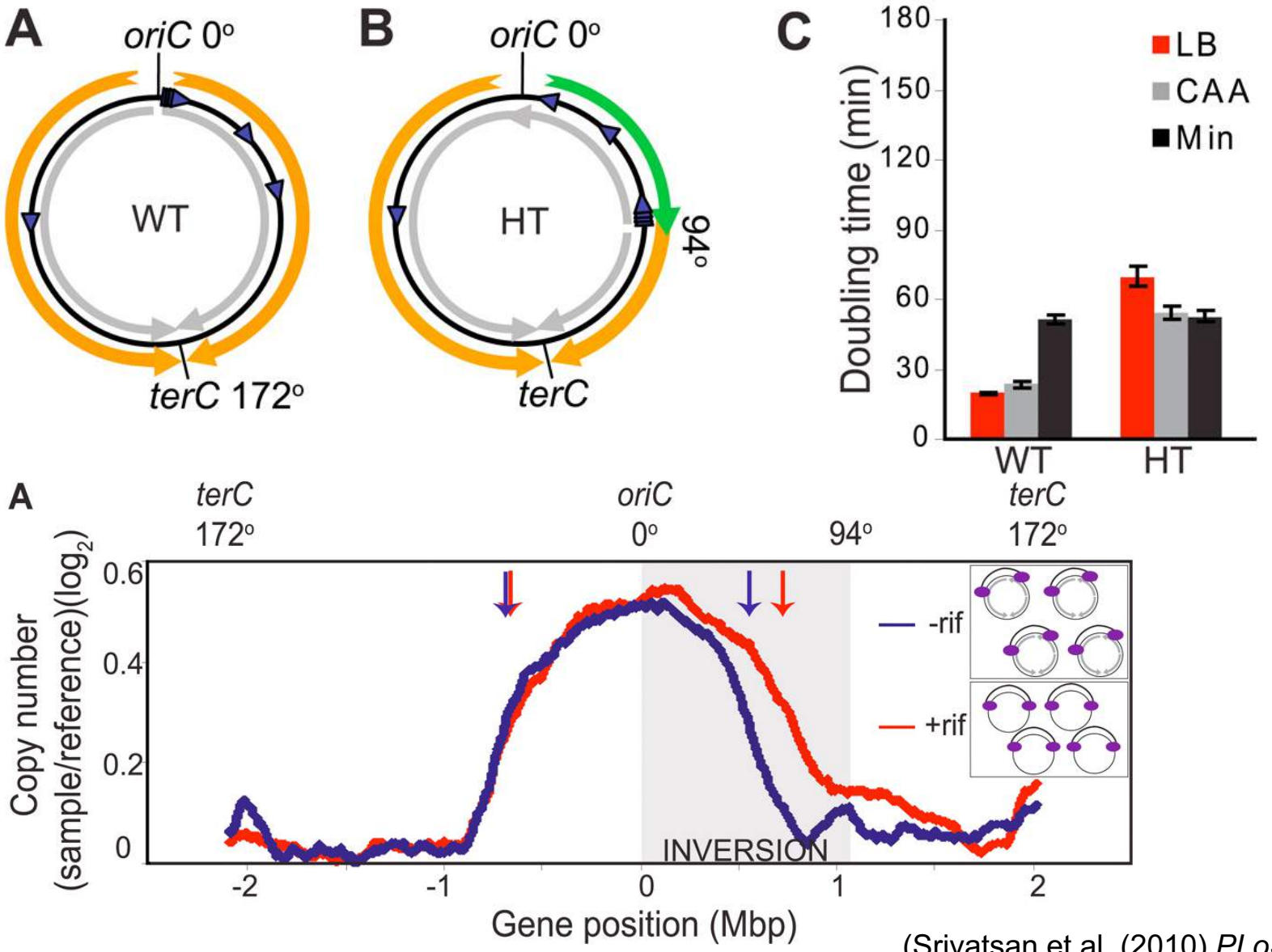
Replication vs. transcription: the origin of the evolutionary pressure



- a replication fork on the “leading strand” will disrupt the transcription in the operon, but that can restart as soon as the fork passes through the transcriptional origin
- in the case of the “lagging strand” because of the opposite orientation this will take much longer, which could be important for highly transcribed genes

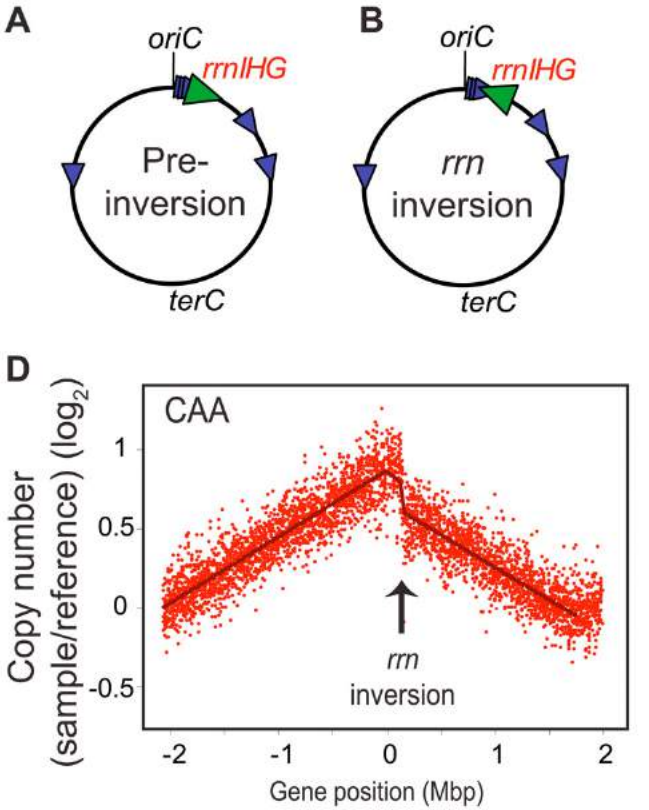
(Price at al. (2005) *Nucleic Acids Res*)

The conflict between replication and transcription ill lead to slower growth



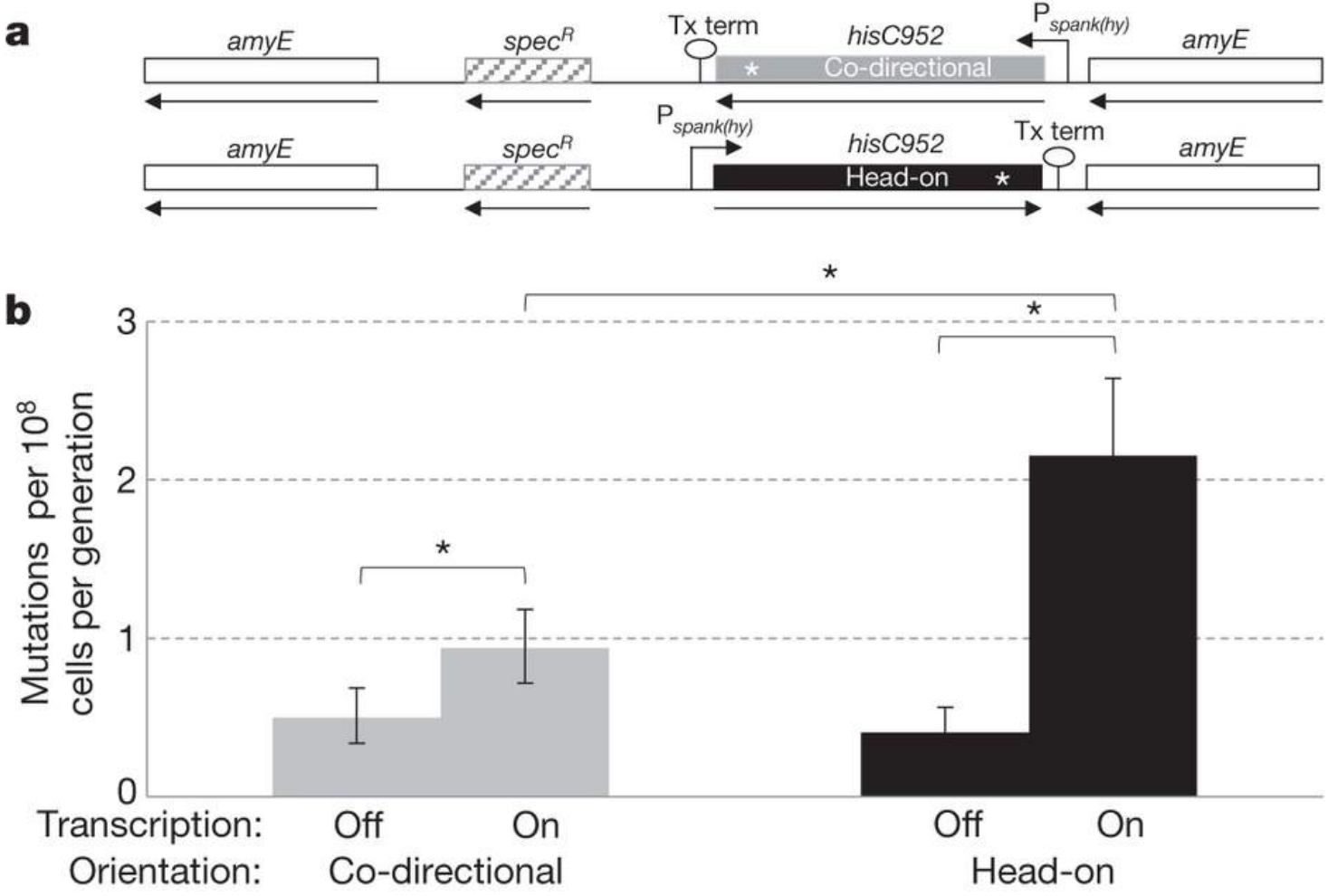
(Srivatsan et al. (2010) *PLoS Genet*)

The reverse orientation of highly transcribed genes will turn on DNA repair



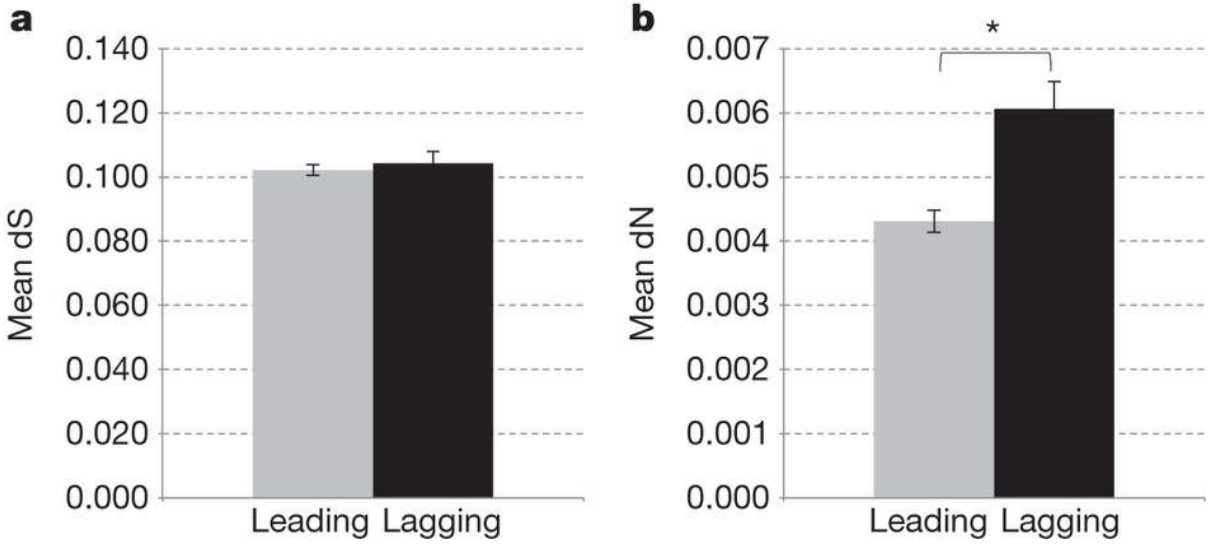
(Srivatsan et al. (2010) *PLoS Genet*)

Opposite orientation leads to higher mutation rates

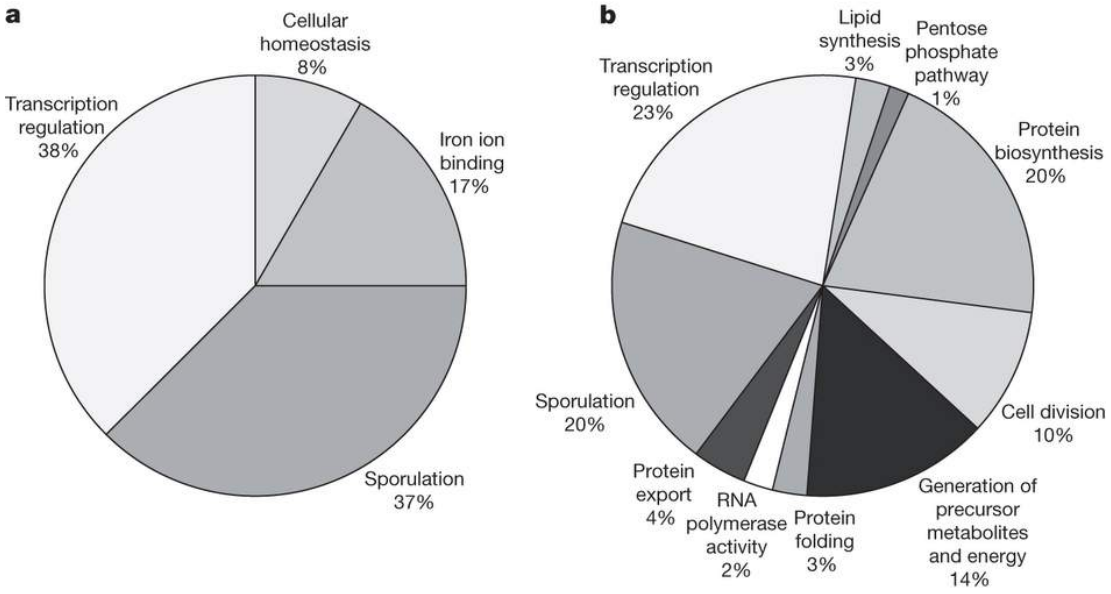


(Paul et al. (2013) *Nature*)

The higher mutation rate of the lagging strand could be an evolutionary advantage for fast changing genes



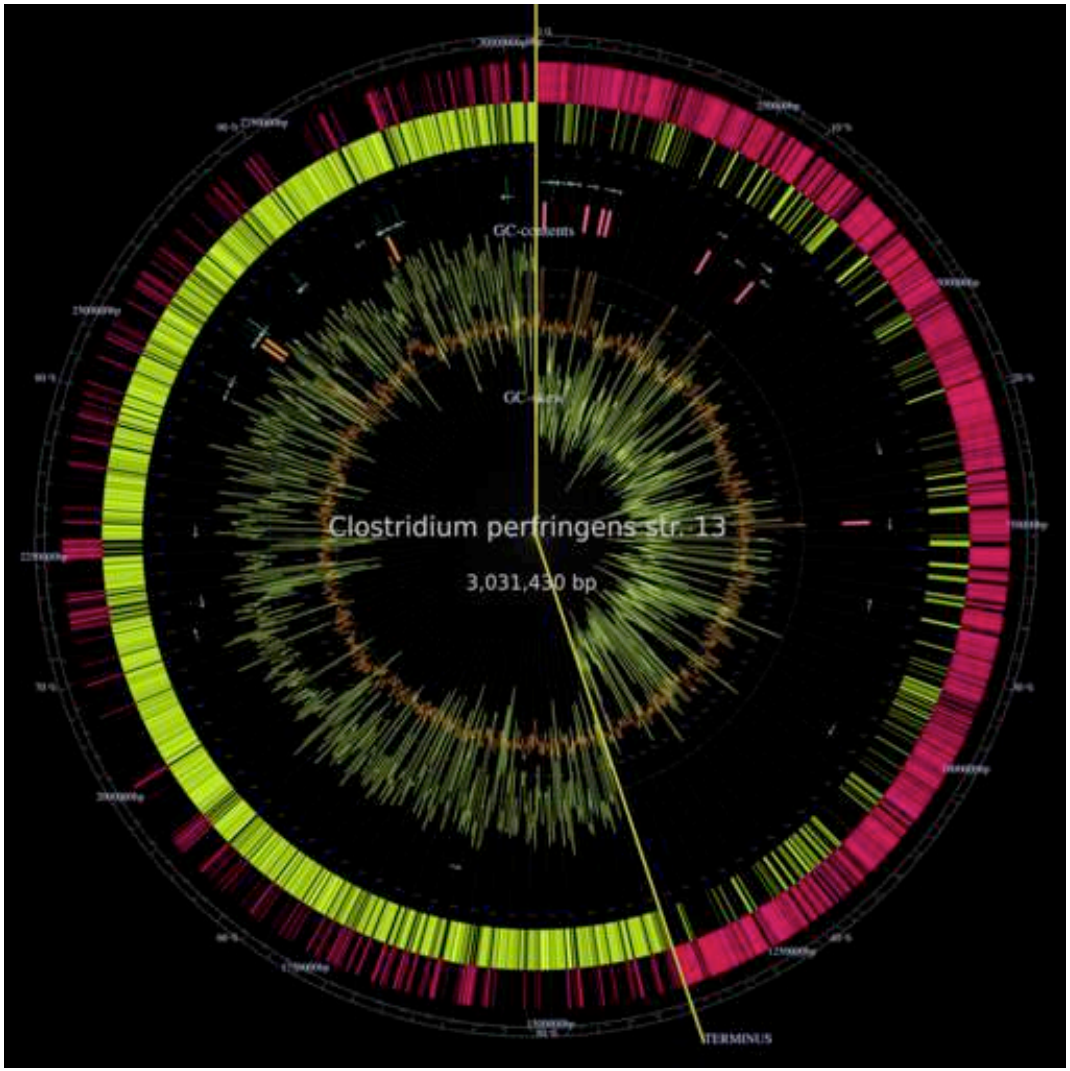
lagging strand



leading strand

(Paul et al. (2013) *Nature*)

“GC-skew” in the genome of *Clostridium perfringens*



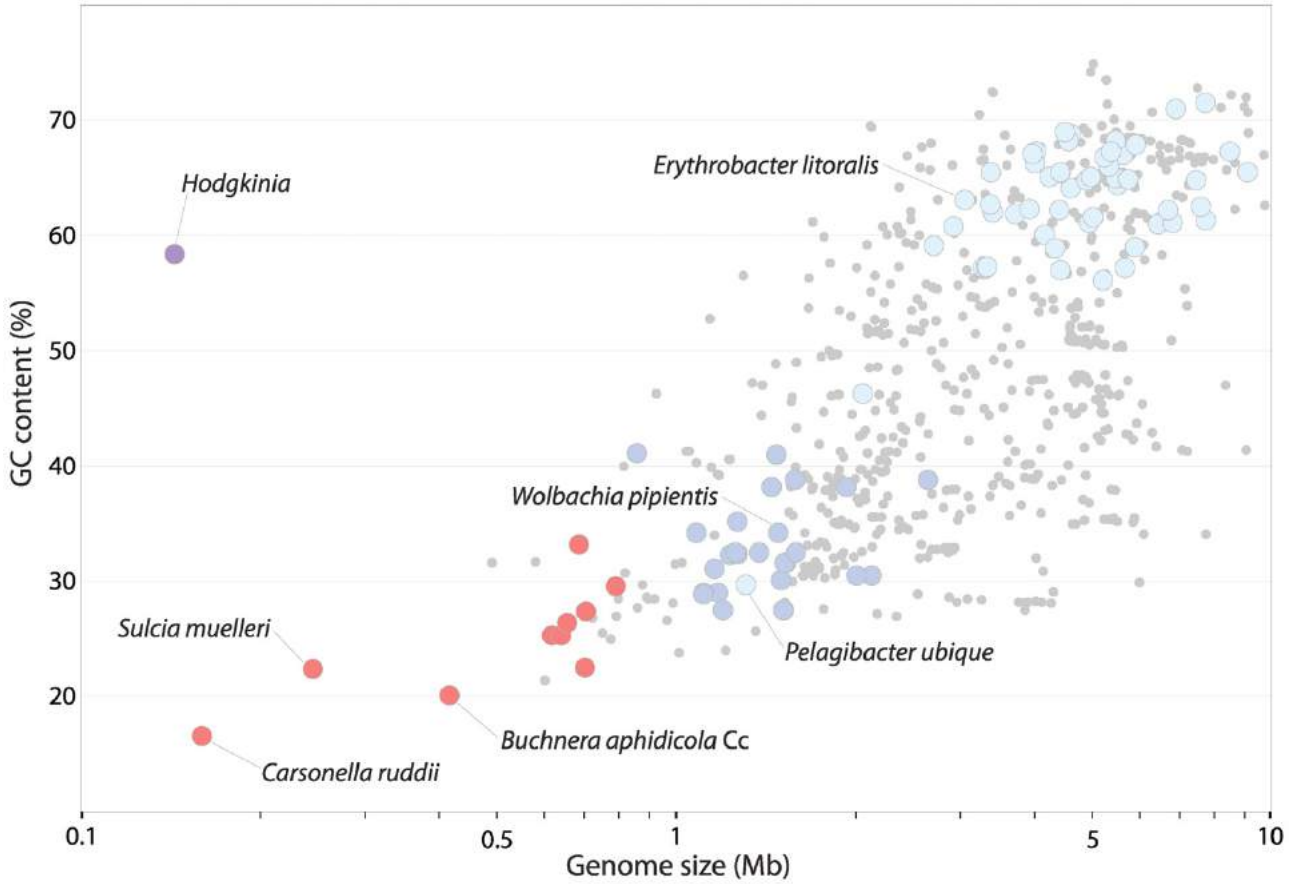
-“GC skew” describes the relative excess of: (C-G)/(C+G).

- because in the bacterial genomes replication prefers G in the leading strand, GC-skew can reveal the replication origin and terminus in the genome

- the *Clostridium* genome is an extreme example of this

(Shimizu et al. (2002) *PNAS*)

GC content vs. genome size

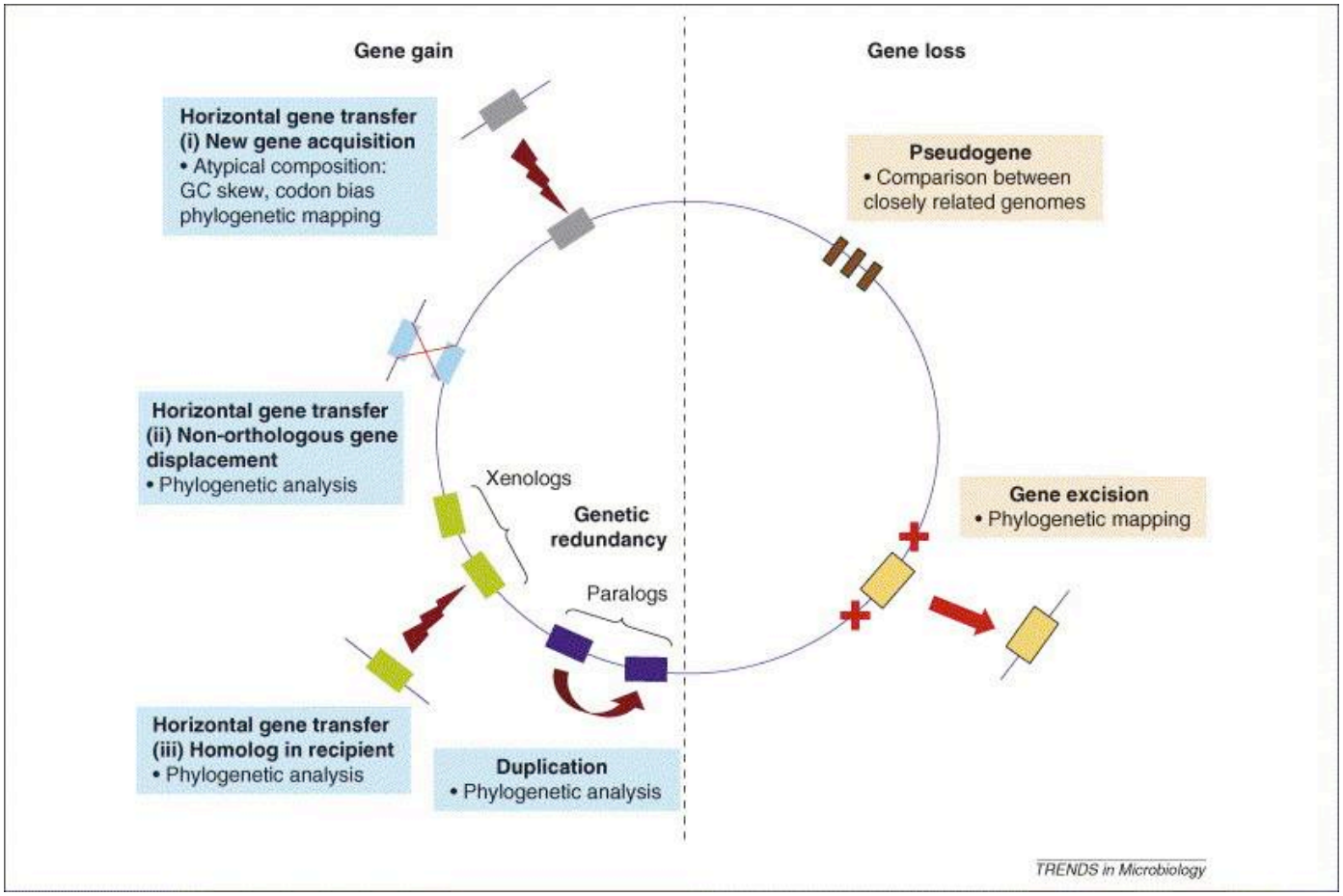


(McCutcheon et al. (2009) *PLoS Genet*)

Two potential explanations:

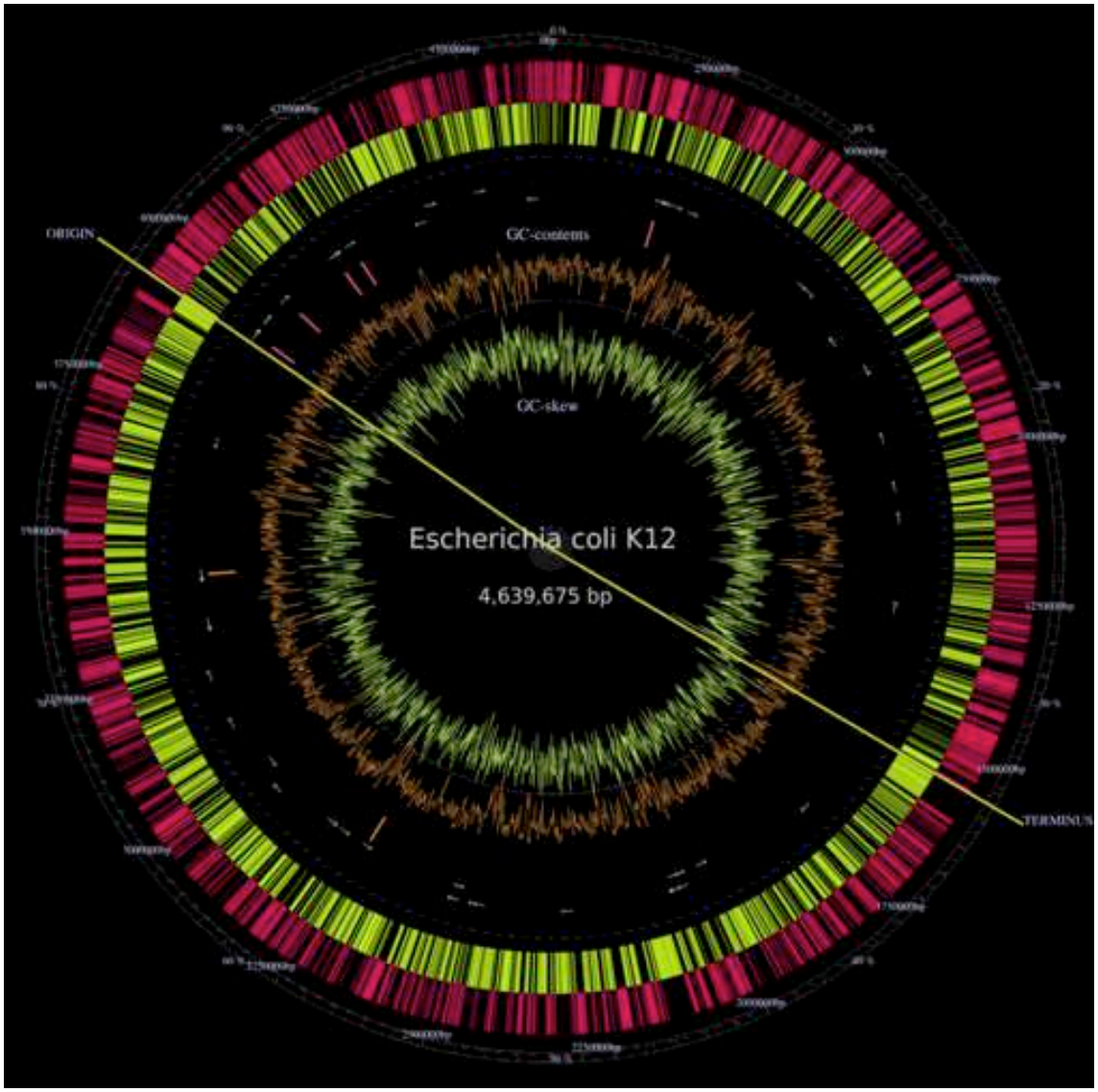
- **energetic reasons**: the synthesis of GTP and CTP requires more energy, and the parasites with small genomes are optimizing for this as well
- **mutation-related reasons:** prokaryotes with small genomes often lost their DNA repair enzymes, and the most common mutation is the C -> T transition

The dynamics of the bacterial genome



(Abby and Daubin (2007) *TiM*)

Eschrichia coli K12

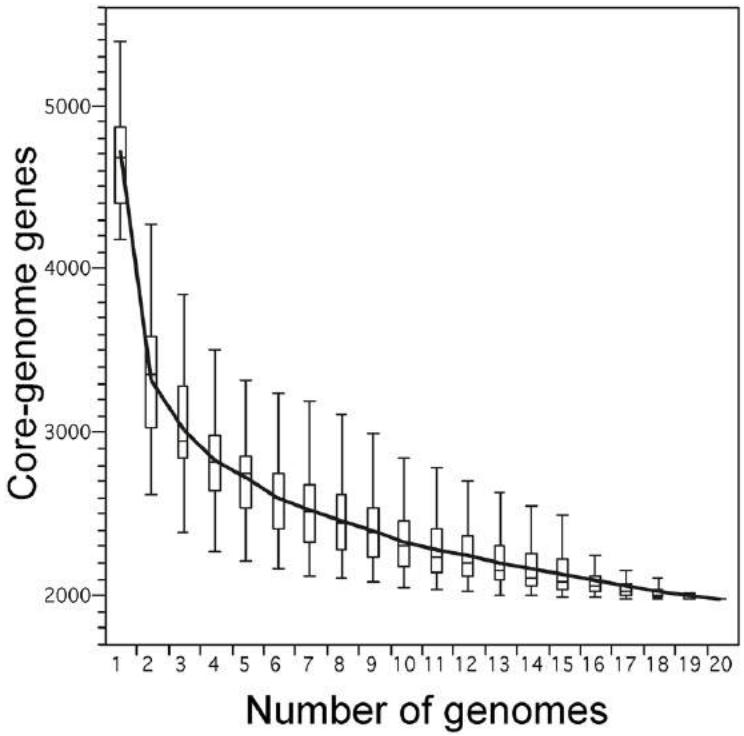


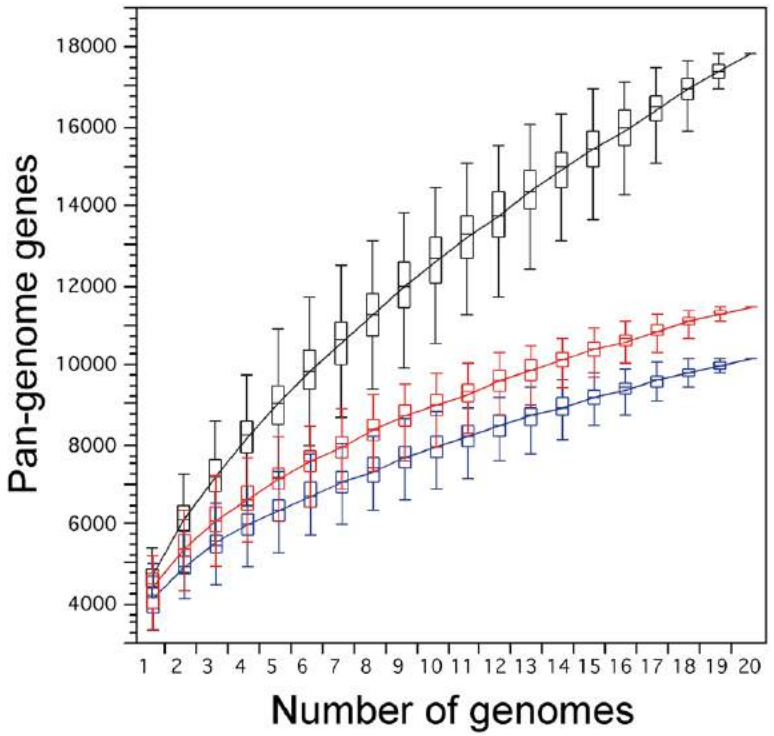
- 4.6 Mb

- 4288 protein coding genes (we still do not understand the function of 1/3 of these)

(Blattner et al. (1997) *Science*)

Eschrichia coli - pan-genome vs core-genome



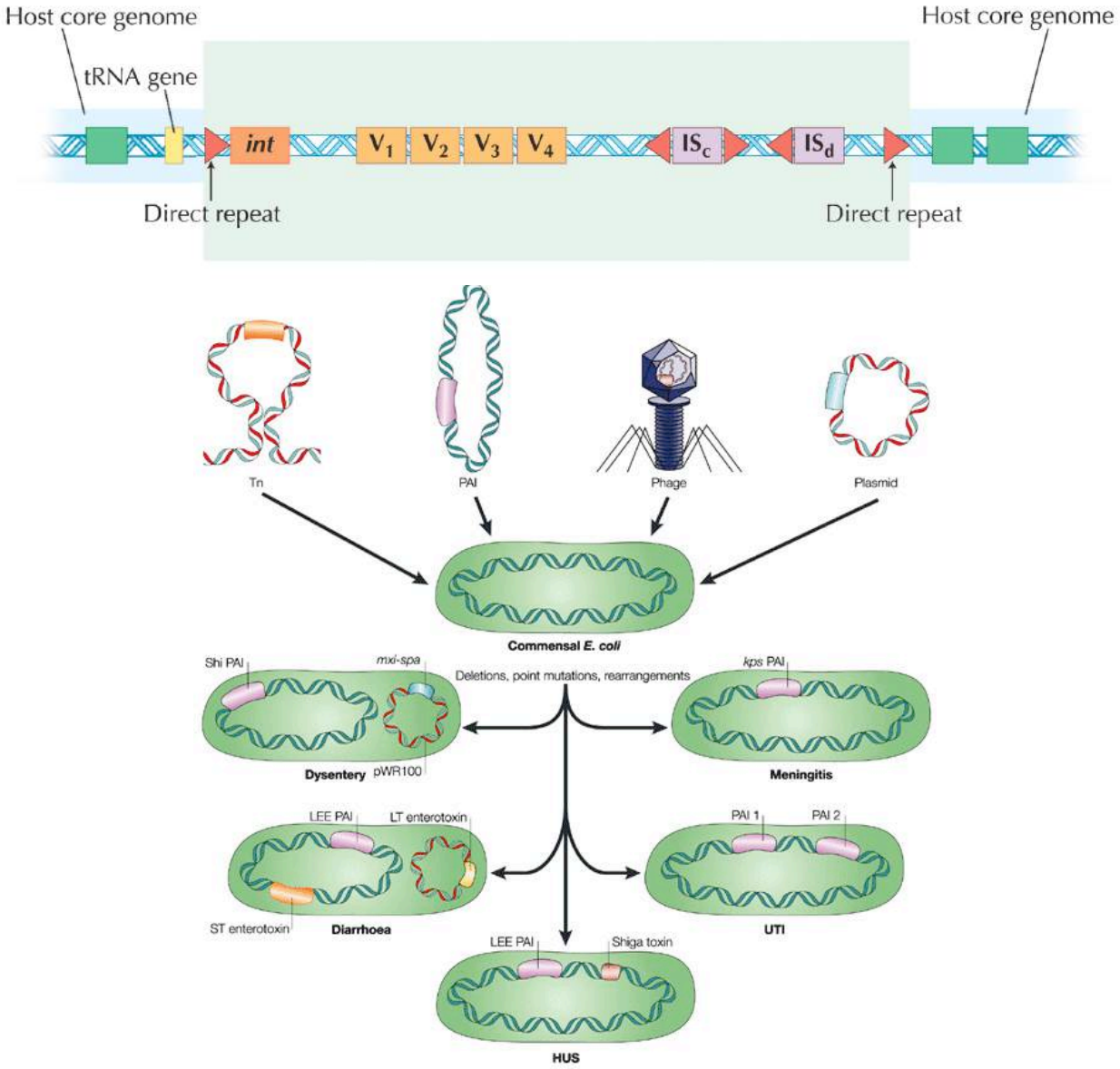


- after sequencing multiple *E. coli* genomes it became obvious that the really essential genes are only a fraction of those found in K12

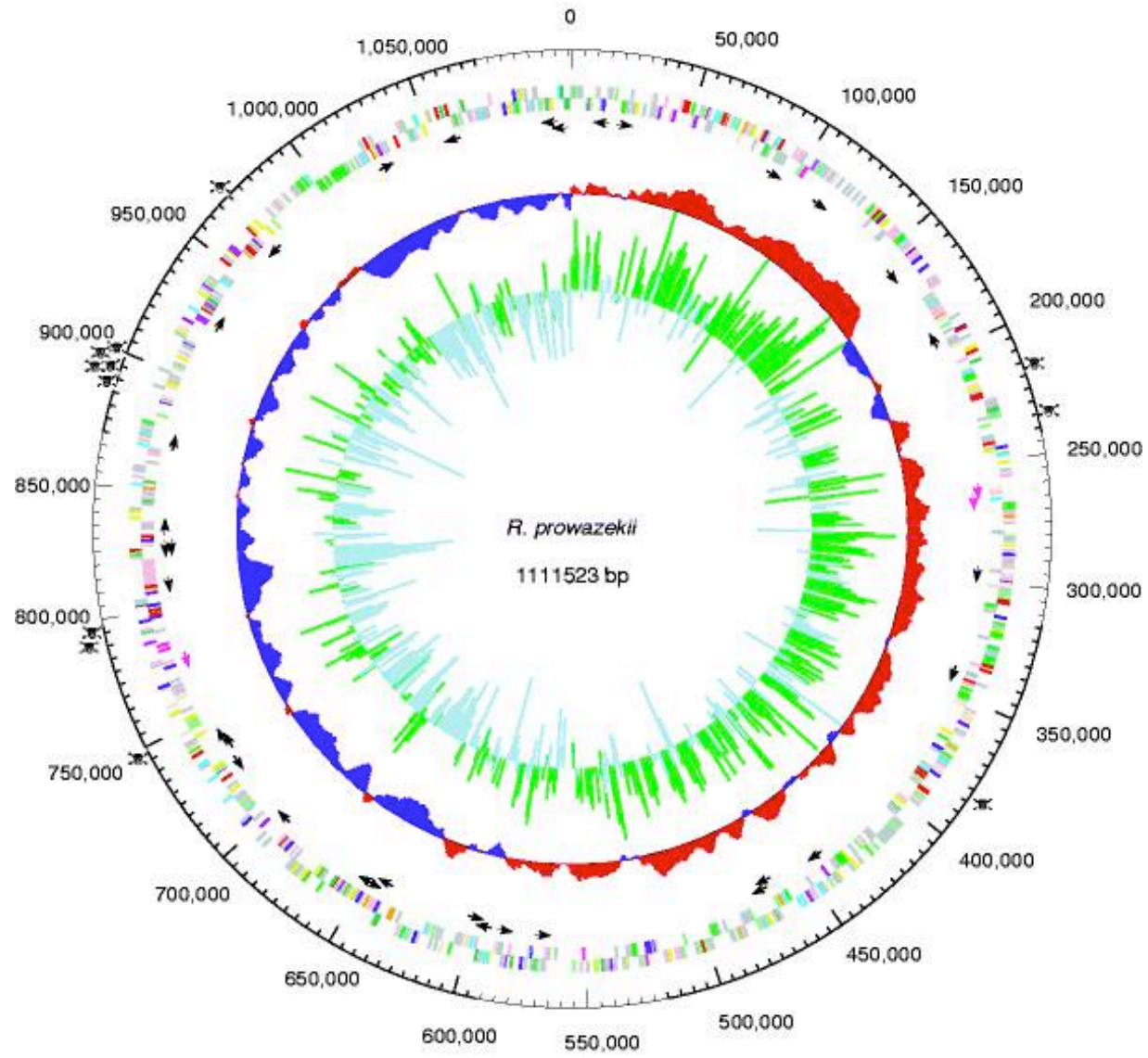
- recent estimates put the size of the *E. coli* core genome to <1900 genes, whereas the pan genome (all the genes that have been found in any *E. coli* isolate) is over 17 000.

(Touchon et al. (2009) *PLoS Genet*)

Pathogenicity islands



Rickettsia prowazekii



(Andersson et al. (1998) *Nature*)

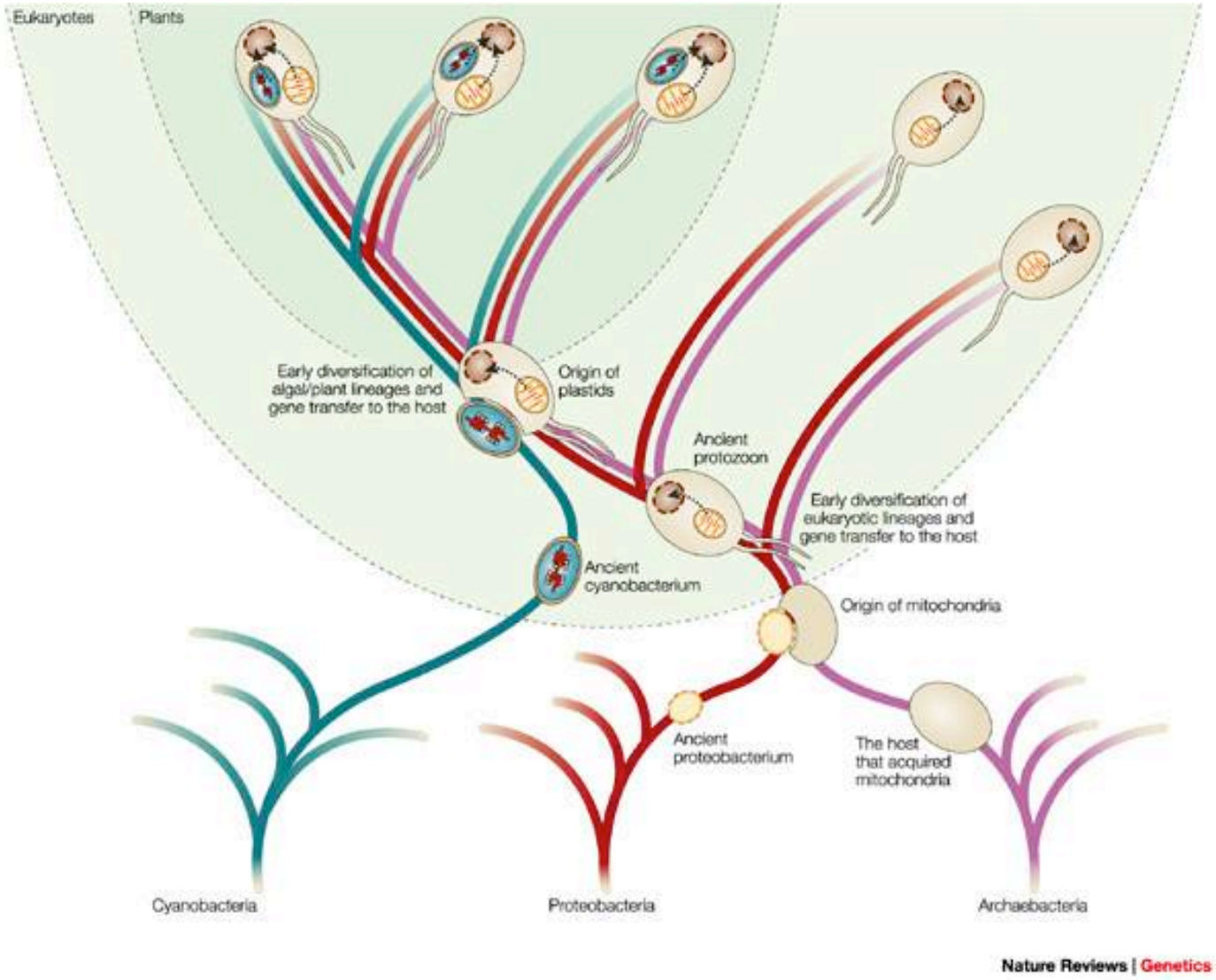
- obligate, intracellular parasite that causes typhus

- a big part of the genome (24%) contains non coding sequences

- these are pseudogenes that gradually acquire more and more mutations

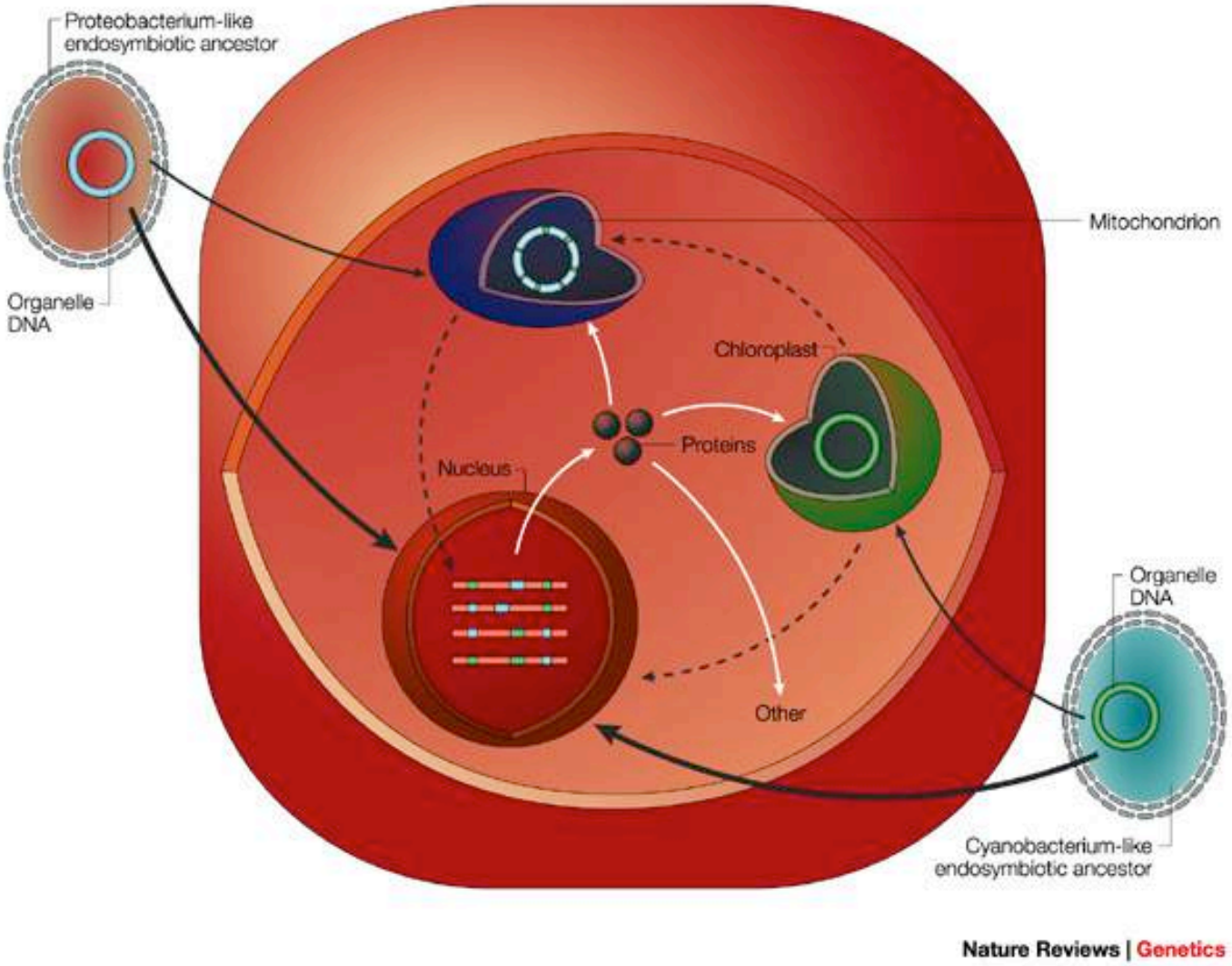
- *Rickettisa* is part of the α-proteobacteria, just like the ancestor of the mictochondrium, therefore its genomic degradation can be informative to understand the evolution of the mitochondrial genome.

Endosymbiotic gene transfers



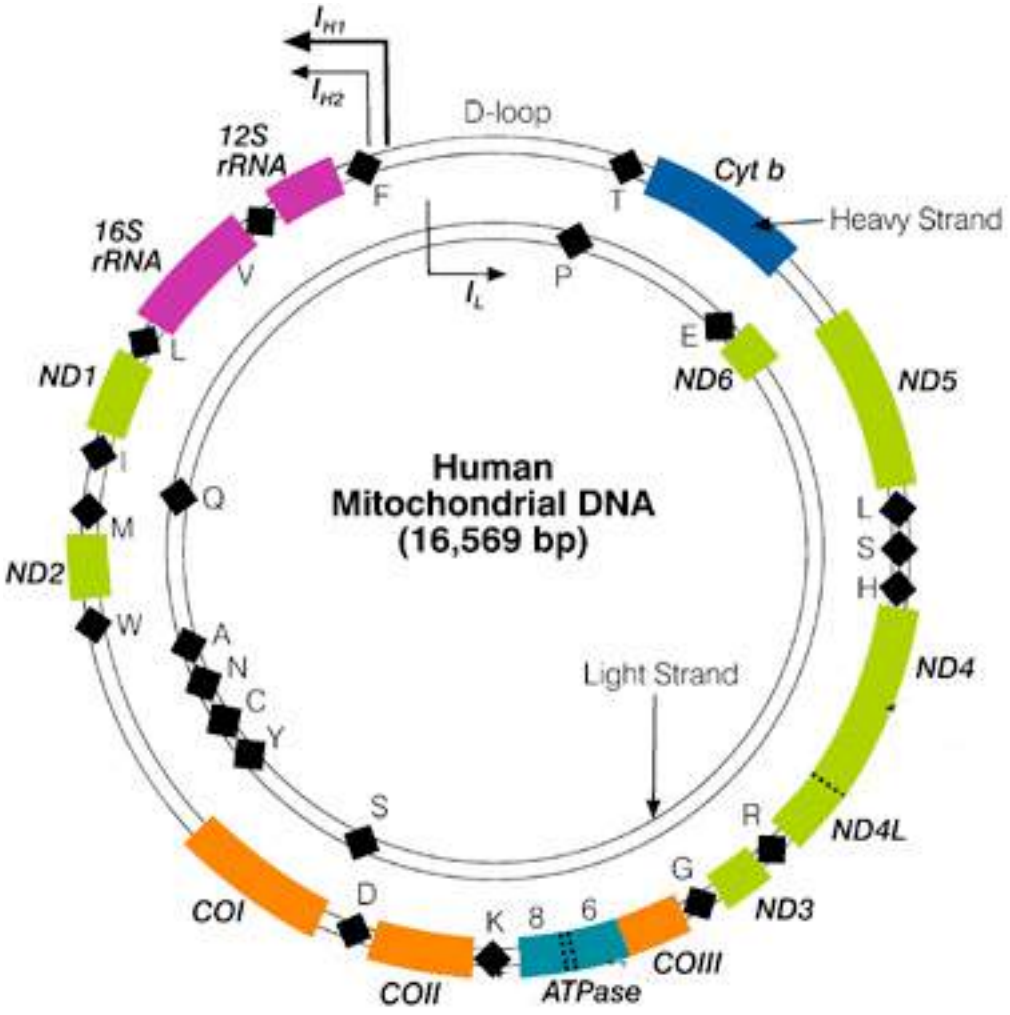
(Timmis et al. (2004) *Nat Rev Gen*)

Endosymbiotic gene transfers



(Timmis et al. (2004) *Nat Rev Gen*)

The mitochondrial genome



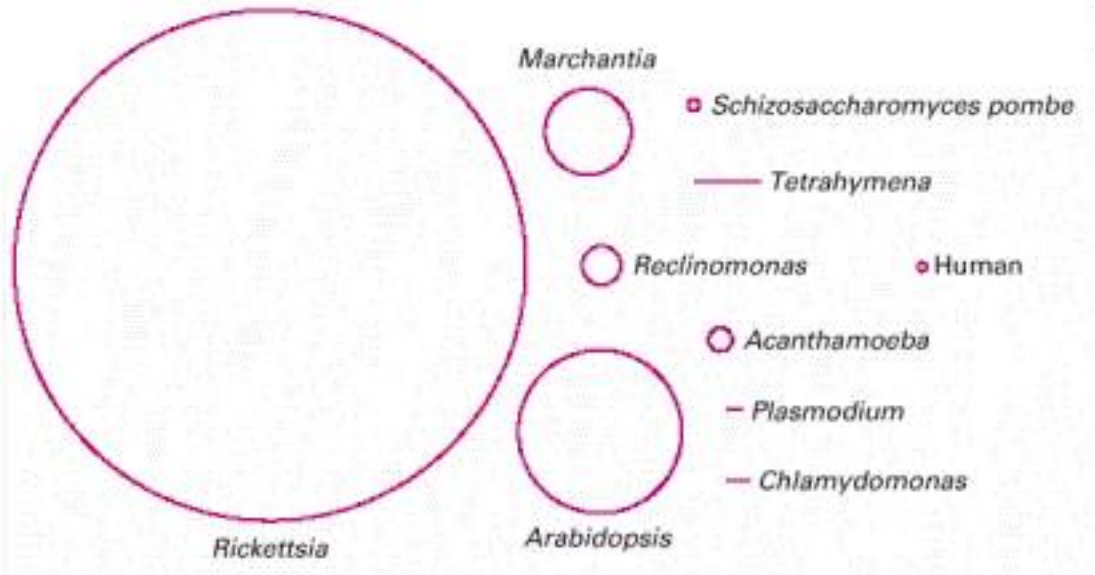
Double stranded, circular DNA, coding on both strands

In humans it is 16,569 bp long

Encodes for 37 genes
13 of these genes are involved in OxPhos, the rest are tRNAs and rRNAs

There are multiple copies in every mitochondrial matrix

Mitochondrial genome size in different eukaryotes



- the size of the mitochondrial genome can be as small as 6000 bp (*Plasmodium falciparum*) or as big as 300,000 bp (some plants)

- most are circular, but some are linear

- in animals (Eumetazoa) mtDNA size is relatively stable, around 16,500 bp

- the *Rickettsia* genome, used as reference is 1.1 million bps long

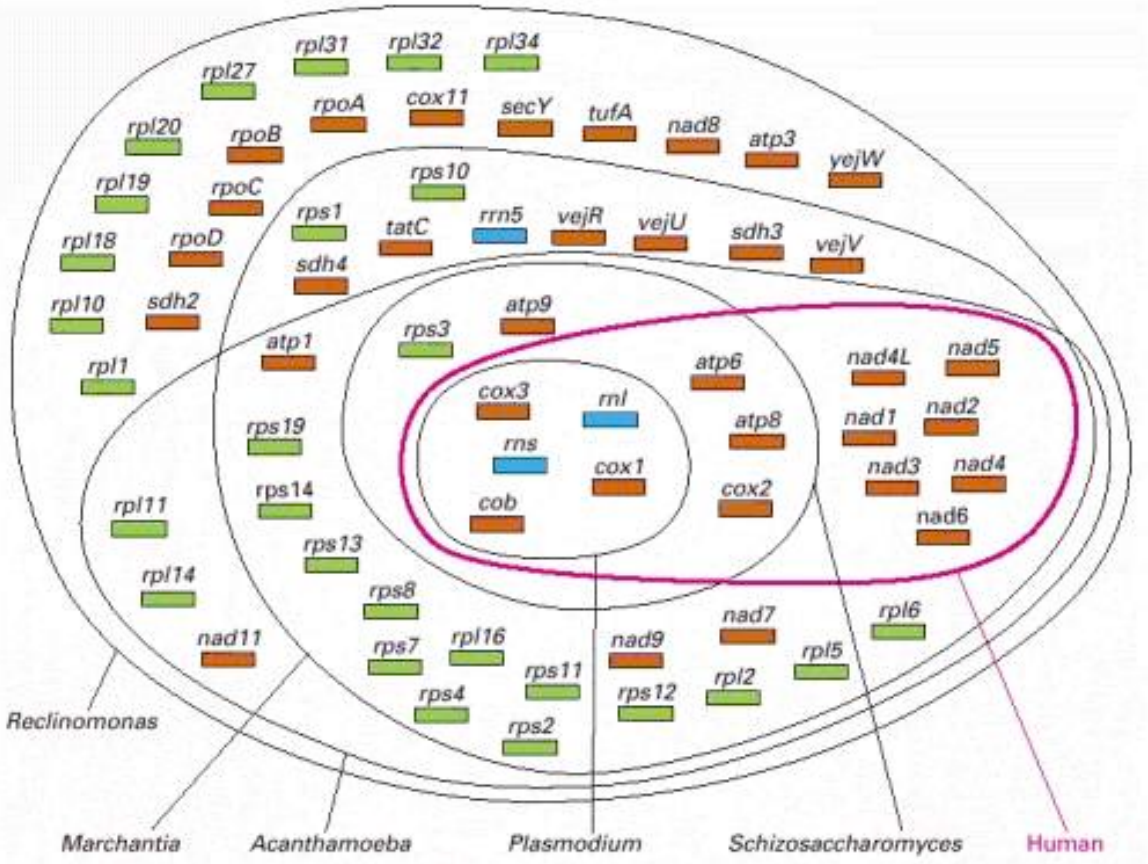
(Alberts et al.: Molecular Biology of the Cell)

Mitochondrial genome size in different eukaryotes

- less complex mitochondrial genomes contain subsets of genes from more complicated mt-genomes.

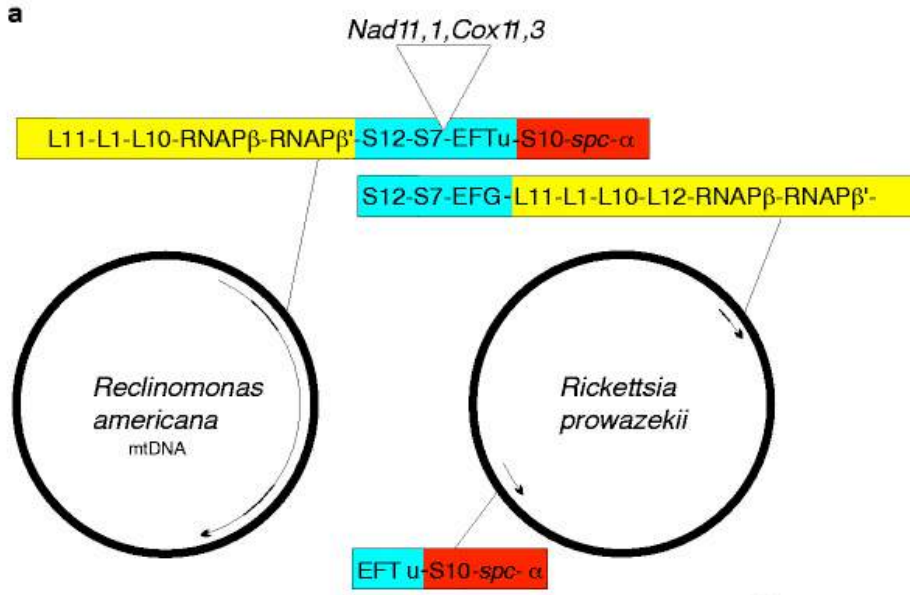
- five genes are present in *all* mtDNAs:: *cob, cox1, cox3, rns, rnl*

- what happened with the other genes…?

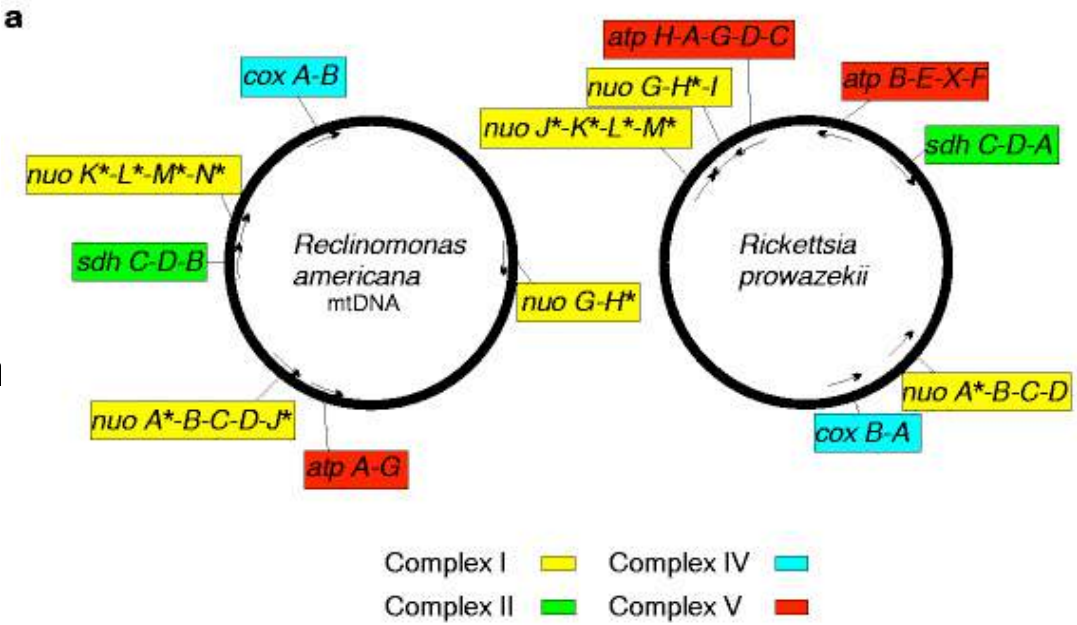


(Alberts et al.: Molecular Biology of the Cell)

The mitochondrial genome and the structure of the *Rickettsia* genome



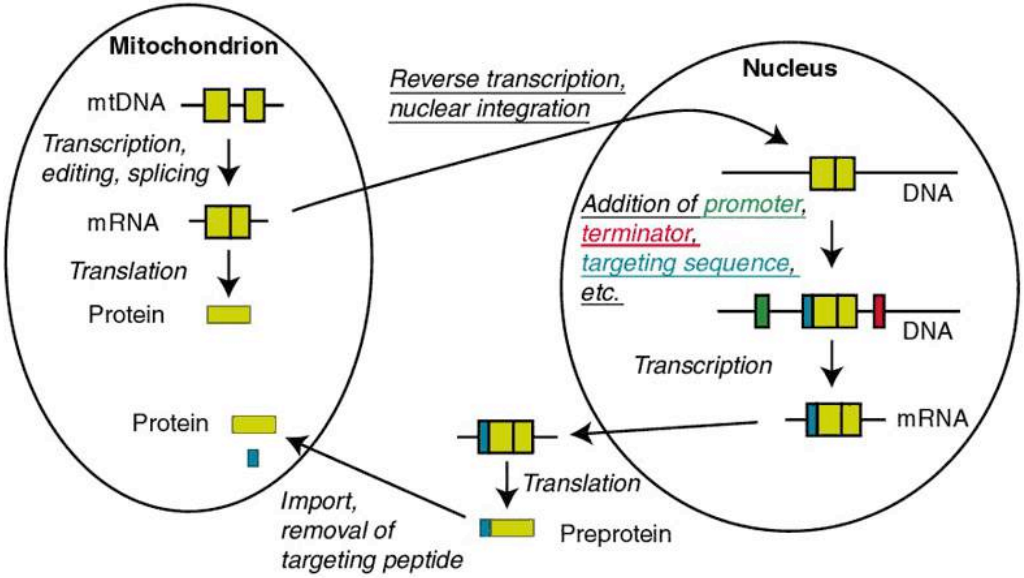
- ribosomal proteins

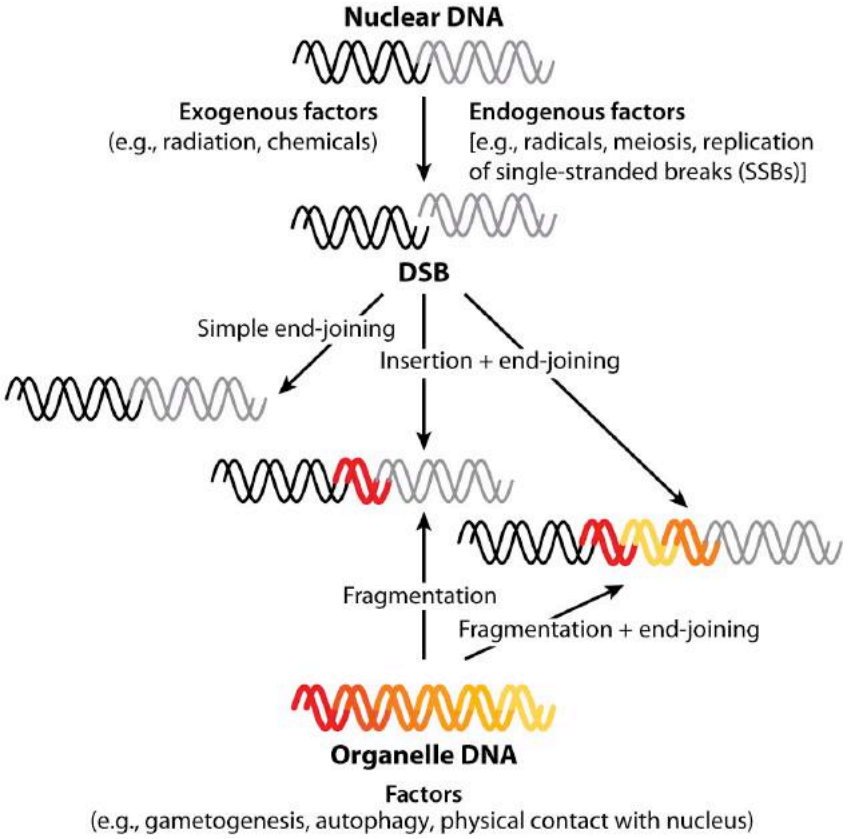


- genes involved in ATP synthesis (ATP synthesis happens similarly in *Rickettsia* and the mitochondrium)

(Andersson et al. (1998) *Nature*)

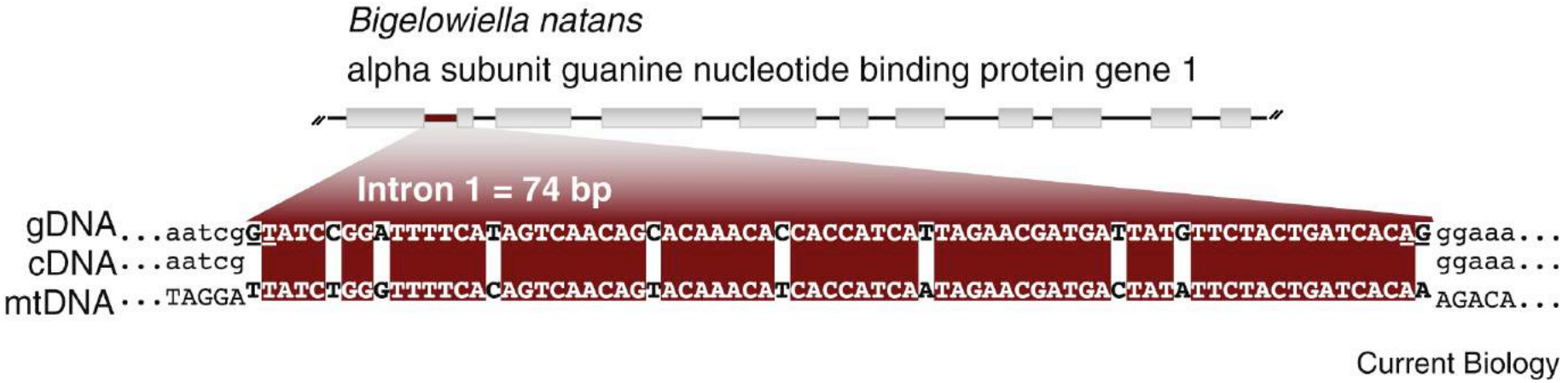
Hypothetical ways for gene transfer





Kleine T, et al. 2009.
Annu. Rev. Plant Biol. 60:115–38

mtDNA as a source for introns

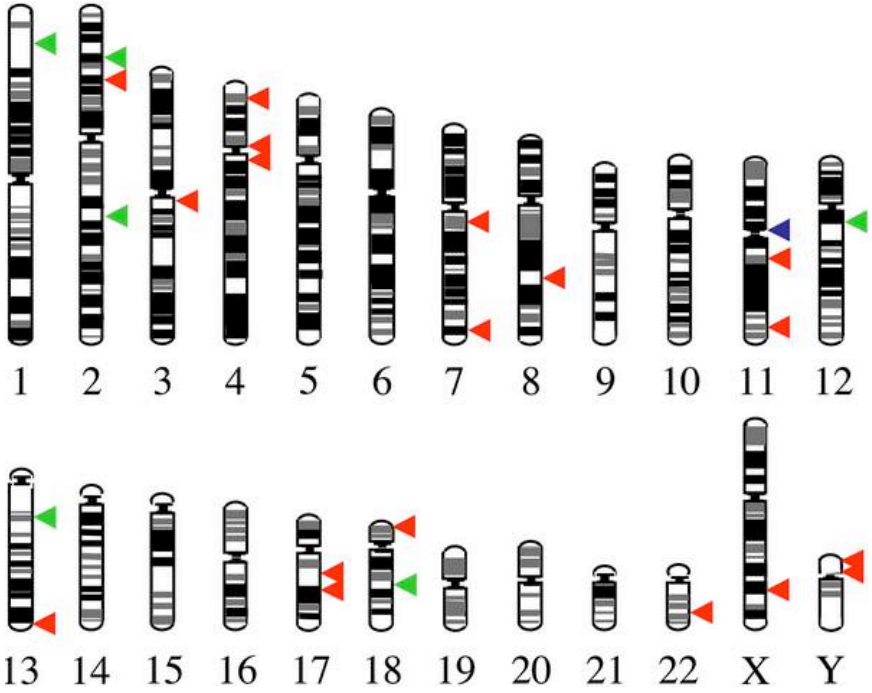


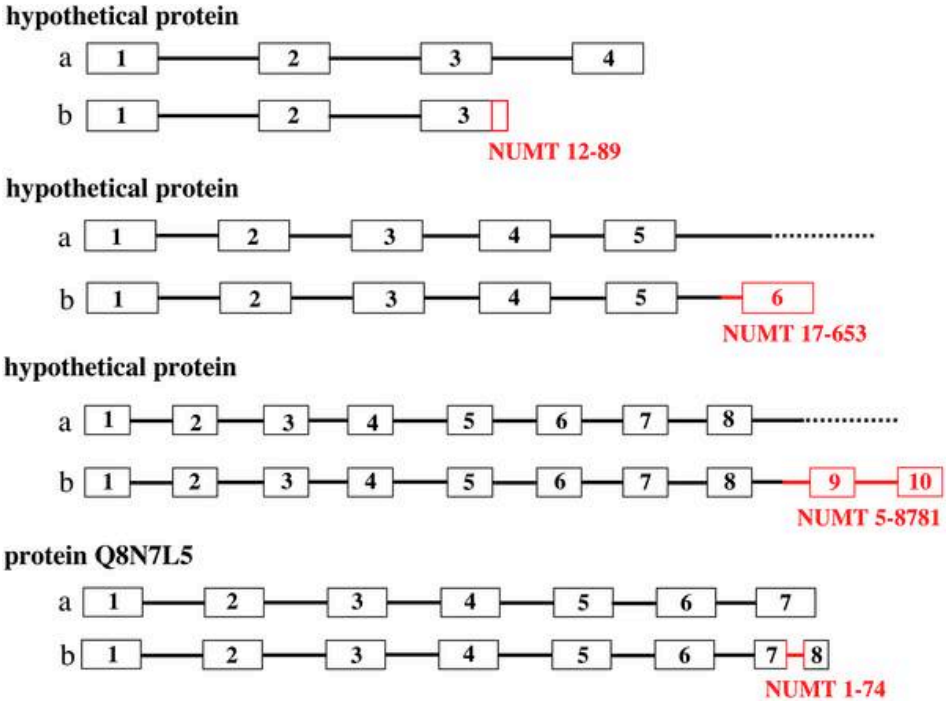
- the first intron of a gene in a unicellular algae is 86% identical to the sequence of the *cox1* mitochondrial gene

- the splice acceptor and donor nucleotides evolved only later – supposedly at the beginning this sequence had suboptimal splicing

(Curtis and Archibald (2010) *Curr Biol*)

mtDNA as a source for introns





- only in the human genome there are 27 specific NUMTs – these arose and got fixed during the past 4-6 million years
- most of them integrated into introns

(Ricchetti et al. (2004) *PLoS Biol*)

The genetic code of the mtDNA in some phyla is different from the ”universal code”

		MITOCHONDRIAL CODES			
CODON	"UNIVERSAL" CODE	MAMMALS	INVERTEBRATES	YEASTS	PLANTS
UGA	STOP	*Trp*	*Trp*	*Trp*	STOP
AUA	Ile	*Met*	*Met*	*Met*	Ile
CUA	Leu	Leu	Leu	*Thr*	Leu
AGA, AGG	Arg	*STOP*	*Ser*	Arg	Arg

*Italics and color shading indicate that the code differs from the "Universal" code.

- in plants and *Reclinomonas* species with large mtDNA genomes the mtDNA code is “universal”

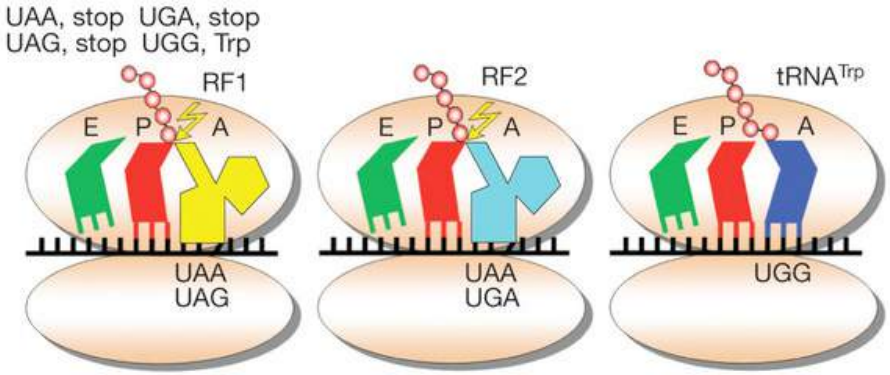
- the STOP -> Trp change can also be observed in some symbiotic/parasitic bacteria

- due to the low number of coding sequences the mitochondrial genome could be more tolerant for changing some rare codons

(Alberts et al.: Molecular Biology of the Cell)

An alternative genetic code in a parasitic bacteria

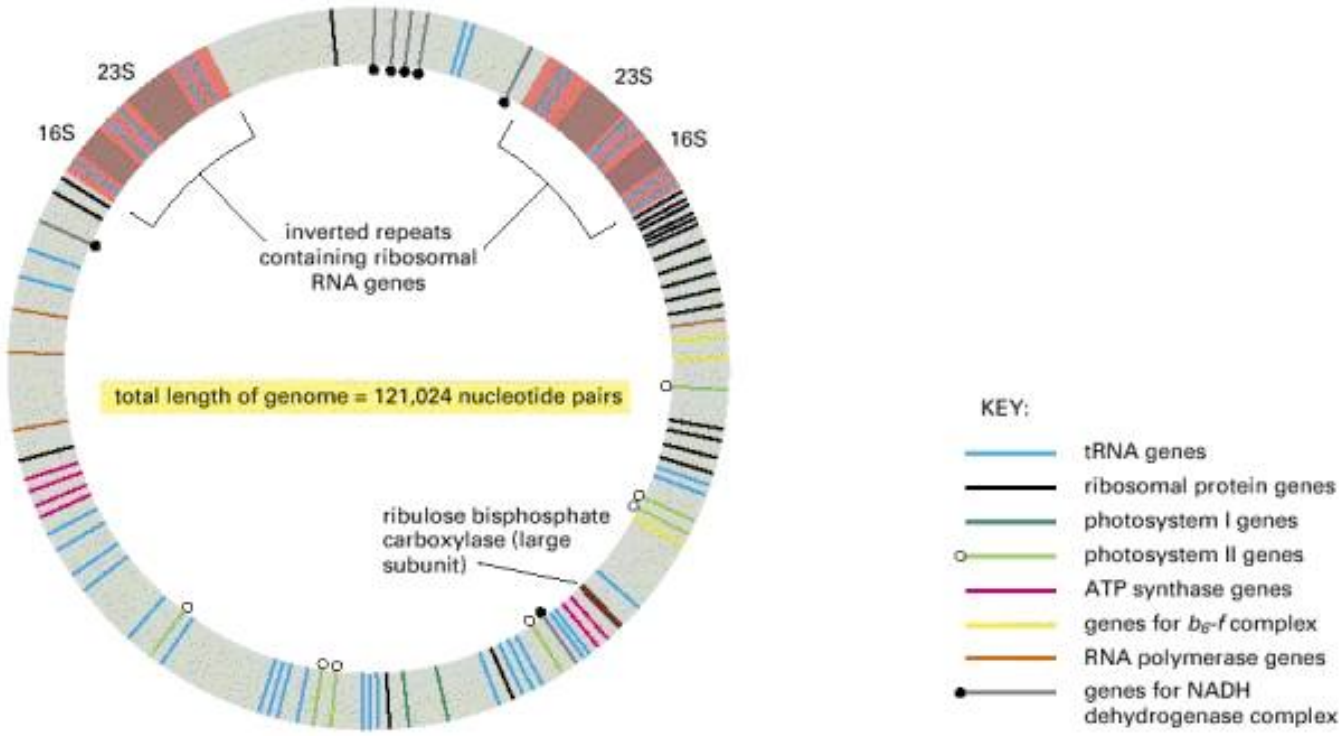
	DnaE (335)	RpoB (711)	RpoC (131)
Hodgkinia	SDFTL.AKAHN	VAFMC.NGFNY	PVVHA.FHGSA
Mloti	ADFIKWAKAQG	VAFMPWNGYNY	PVAHIWFLKSL
Ccres	SDFIKWGKAHG	VAFMPWNGYNF	PVAHIWFLKSL
Pdeni	ADFIKWAKEHN	VAFMPWNGYNY	PVAHIWFLKSL
Rrubr	ADFIQWAKDAD	VAFMPWNGYNF	PVAHIWFMKSL
Elito	ADFIQWAKDHG	VAFMPWNGYNY	PVAHIWFLKSL
Pubiq	SDYIKWAKNND	VAFMPWQGYNF	PVAHIWFLKSL
Rrick	SDFIKWSKKEG	VAFLPWNGYNF	PVAHIWFLKSL
Ecoli	MEFIQWSKDNG	VAFMPWNGYNF	PTAHIWFLKSL
Nmeni	QDFINWAKTHG	IAFMPWNGYNY	PVAHIWFLKSL
Gmeta	ADFINWAKDHG	VAFMPWGGYNF	PVAHIWFLKSL



	tRNA-Trp anticodon	release factors	UGA encodes
initial state	CCA	RF1 RF2	STOP
(1) **mutation of tRNA-Trp gene**			
some readthrough of UGA	*CCA	RF1 RF2	STOP Trp
(2) **loss of Release Factor 2 (RF2)**			
only UAA and UAG read as stop	*CCA	RF1	Trp
(3) **mutation of tRNA-Trp anticodon**			
UGA, UGG both read by wobble rules	UCA	RF1	Trp
(4) **genomic codon adaptation**			
new UAA and UAG stops generated; some UGG codons changed to UGA	UCA	RF1	Trp

(McCutcheon et al. (2009) *PLoS Genet*)

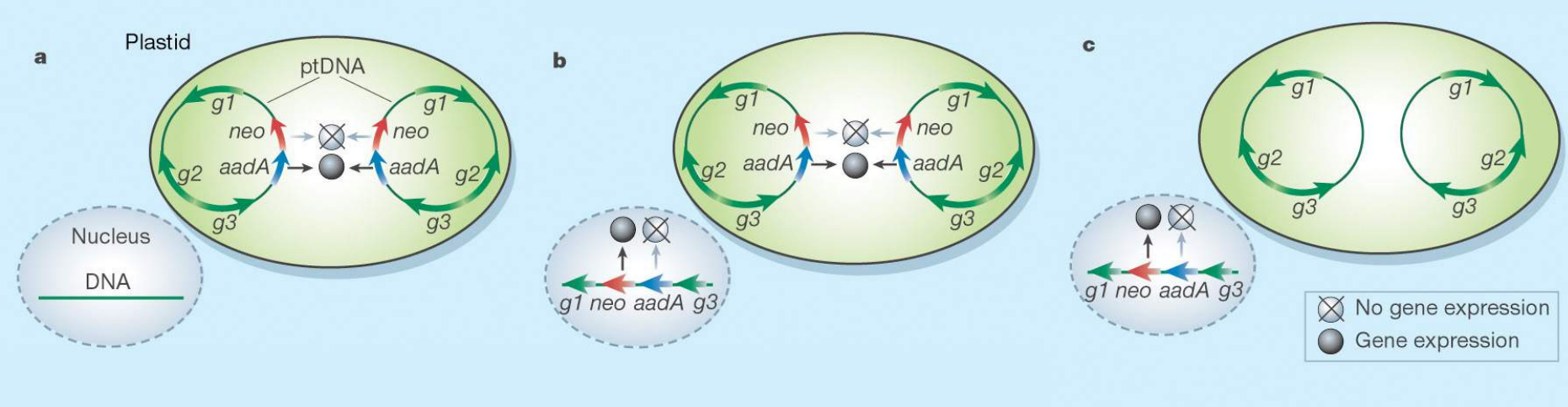
The chloroplast genome



- Double stranded, circular DNA, coding on both strands

- genes regulating transcription are almost identical with their bacterial homologs

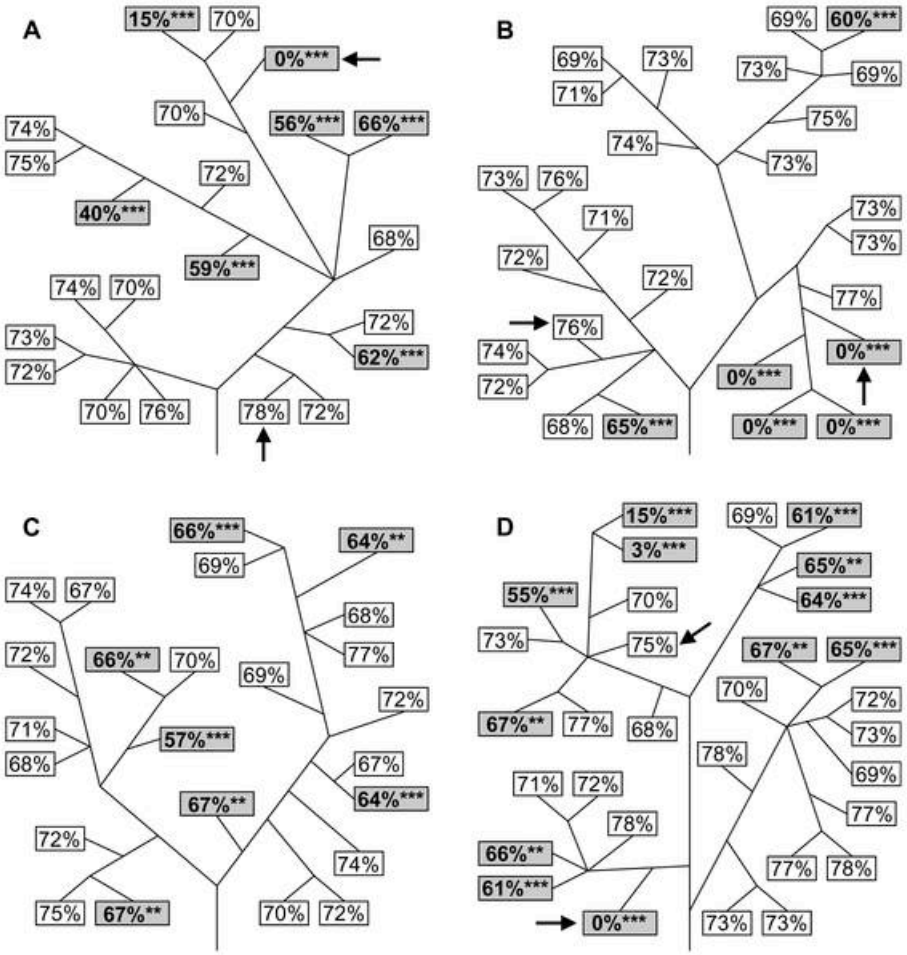
DNA transfer from the chloroplast to the nucleus



- a genetic screen to test for DNA transfer from the chloroplast to the nucleus:
  - the gene encoding for spectinomycin resistence (*aadA*) is behind a bacterial promoter (therefore it is active in the chloroplast), whereas the gene for neomycin resistence *(neo*) is behind a eukaryotic promoter so it is active only in the nucleus
  - in somatic cells (e.g. leaf) the chance of the transfer was 1 : 5 million, whereas in pollen cells 1 : 16 000 (the difference could be due to the fact that during pollen formation the chloroplast breaks down and there is a higher chance for its DNA to get to the nucleus)

(Maliga (2003) *Nature*)

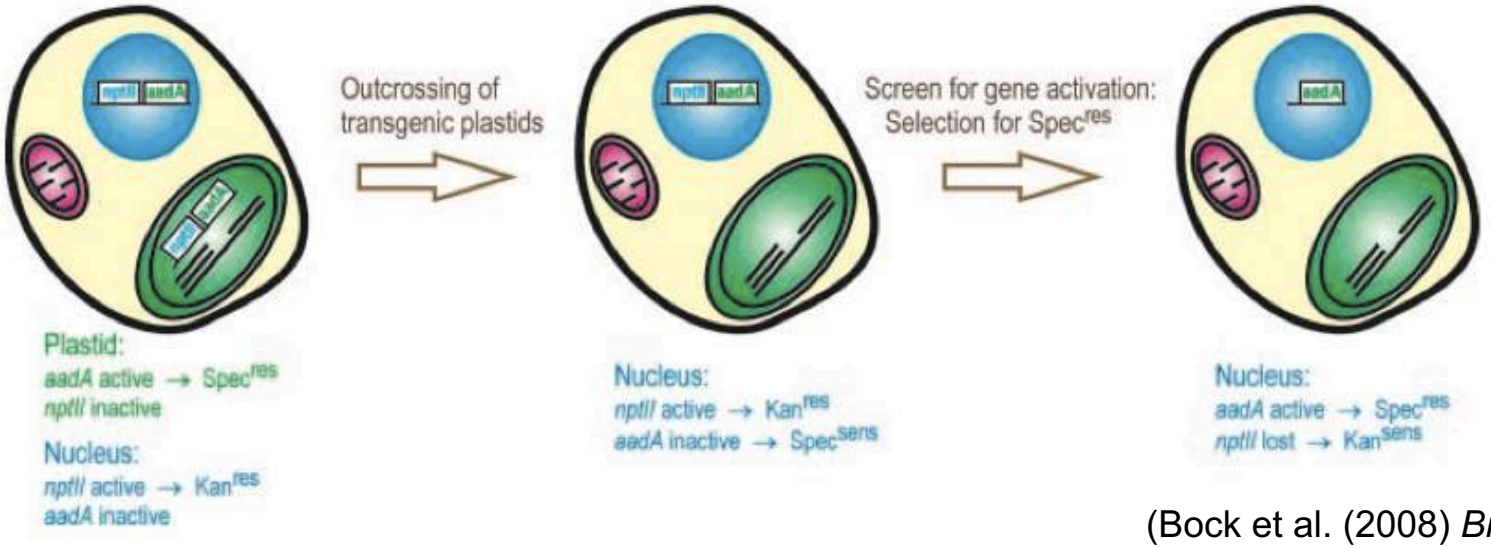
Most of the NUPTs are highly instable



- in self-pollinating systems it can be observed that frequent integrations are counteracted with frequent deletions
- sometimes integrations can get lost within a single generation (the mechanism for this is unknown)

(Sheppard and Timmins (2009) *PLoS Gen*)

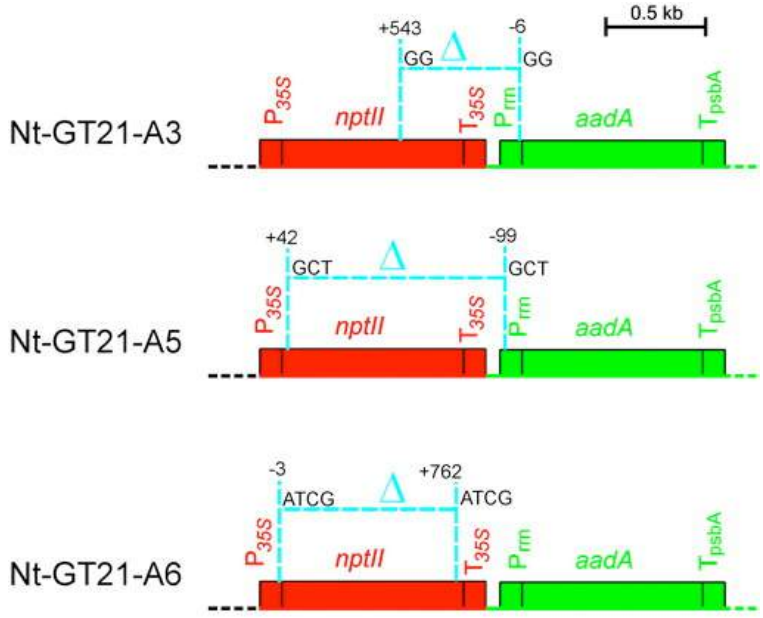
Reactivation of chloroplast genes in the nucleus



(Bock et al. (2008) *Bioessays*)

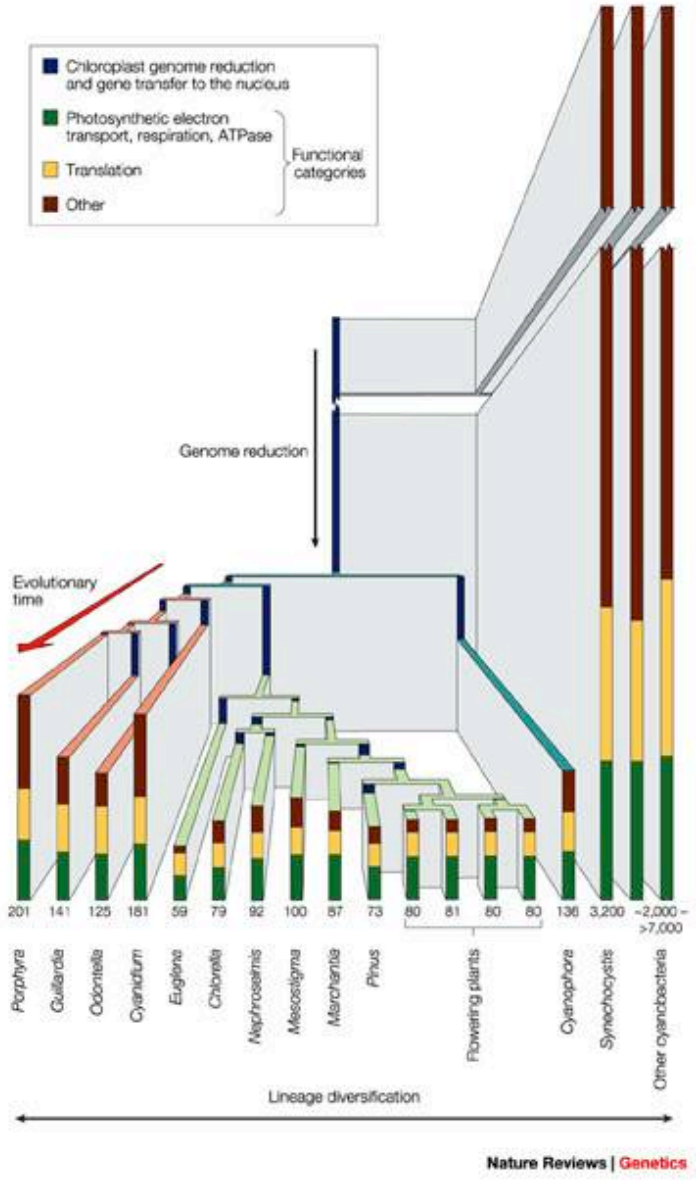
- the frequency of chloroplast-sequence reactivation in the genome is comparable to the frequency of the nuclear transfer

- the activation of nuclear chloroplast sequences happens with the capture of upstream promoters



(Stegemann and Bock (2006) *Plant Cell*)

The evolution of the chloroplast genome



(Timmis et al. (2004) *Nat Rev Gen*)

Why isn't the translocation of mitochondrial and chloroplast genomes to the nucleus complete?

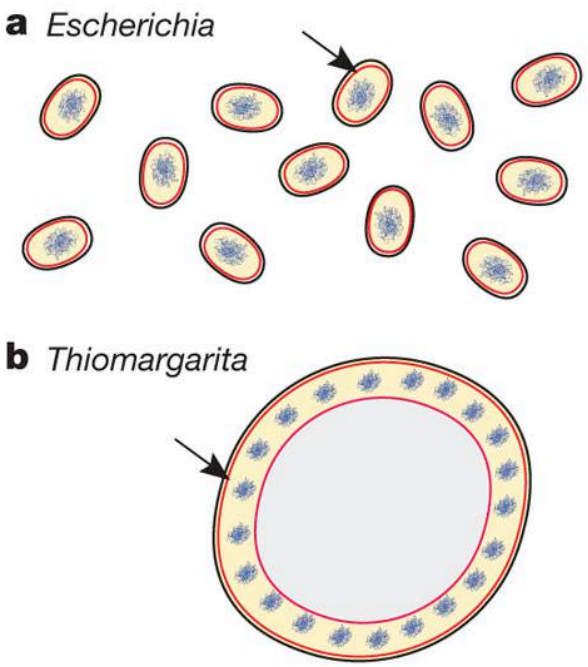
1. Smaller organellar genomes will use alternative codon tables, therefore newly transferred genes will be non-functional in the nucleus.

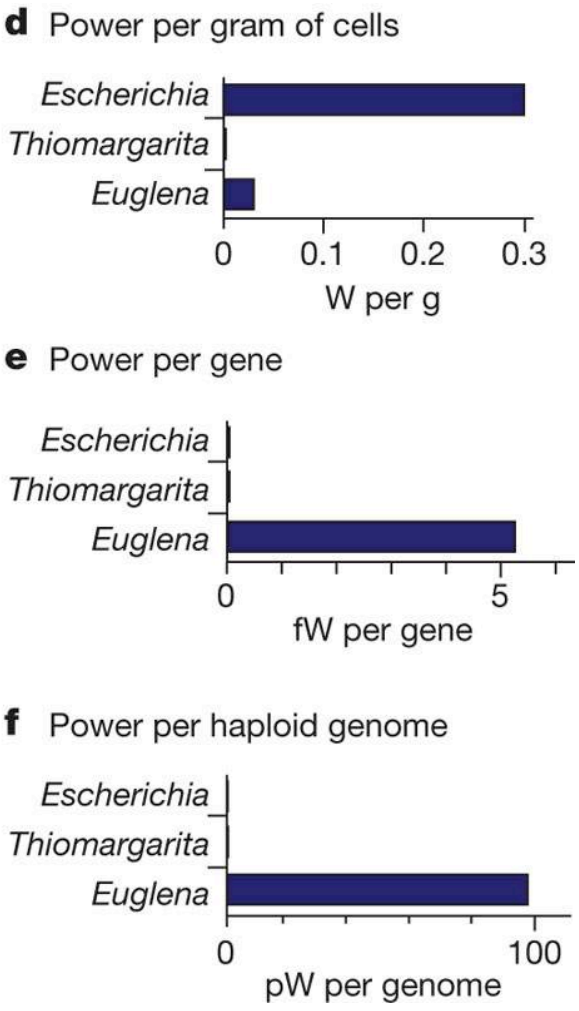
2. Genes encoded by the organellar genome are highly hydrophobic, therefore are hard to transfer through the cytoplasm.

3. CORR *(CO*location of genes and gene-products for *R*edox *R*egulation of gene expression*)* hypothesis: the transcription of some genes is regulated by the redox potential of bioenergetic membranes. These can not be transferred from the organelle, as their regulation is not possible in the nucleus.

(Allen (2003) *Phil Trans R Soc Lond B*)

The complex genome of the eukaryotes was made possible by the emergence of the mitochondrium





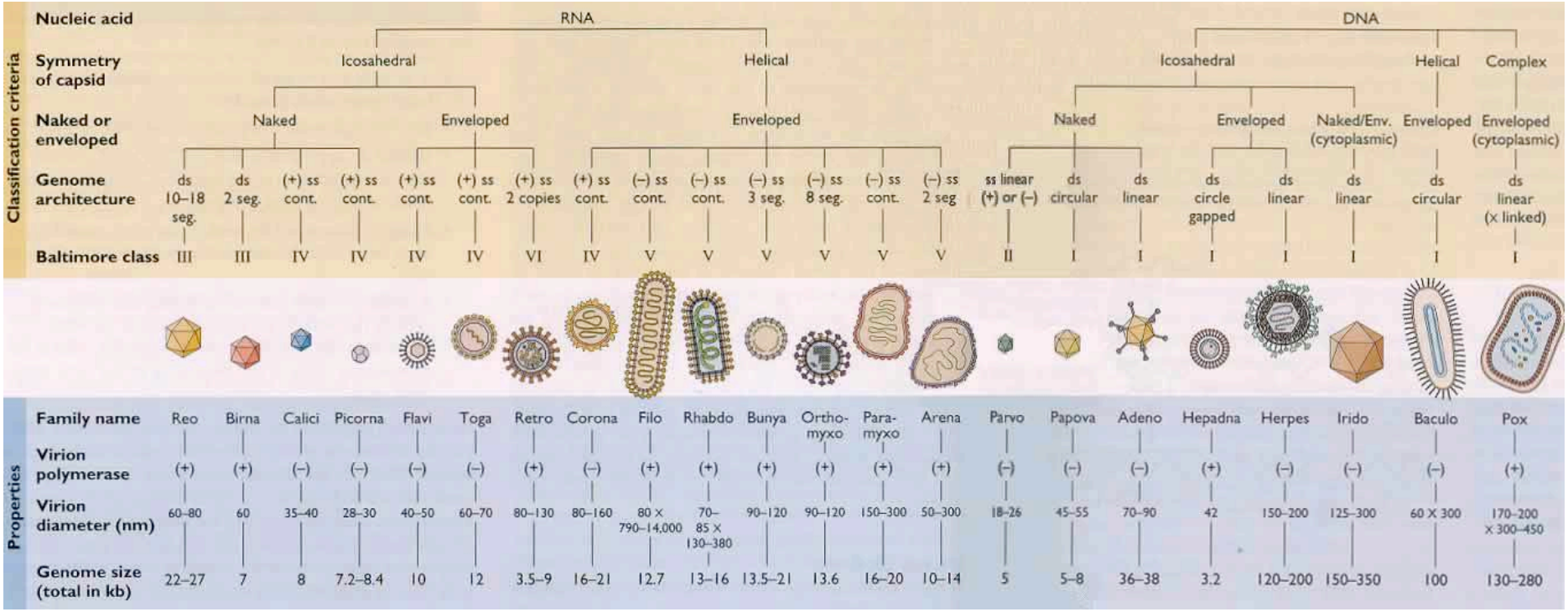
(Lane and Martin (2010) *Nature*)

- the energy production of prokaryotic cells is limited, therefore the energy that can be used to produce one protein depends on the overall protein amount.

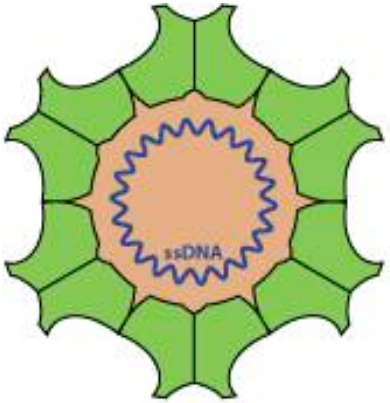
- this establishes an upper limit for the number of possible coding sequences.

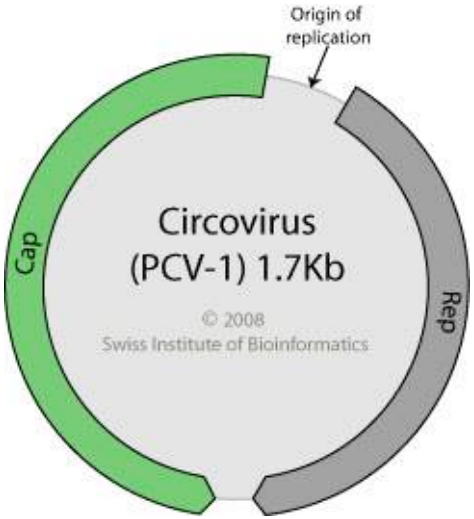
- in eukaryotes with the emergence of the mitochondrium the surface of the bioenergetic membranes increases significantly, ATP can be produced to demand, which results in a 400 000 fold increase in the coding capacity

Viral genome types



The smallest viruses - *Circocoviridae*



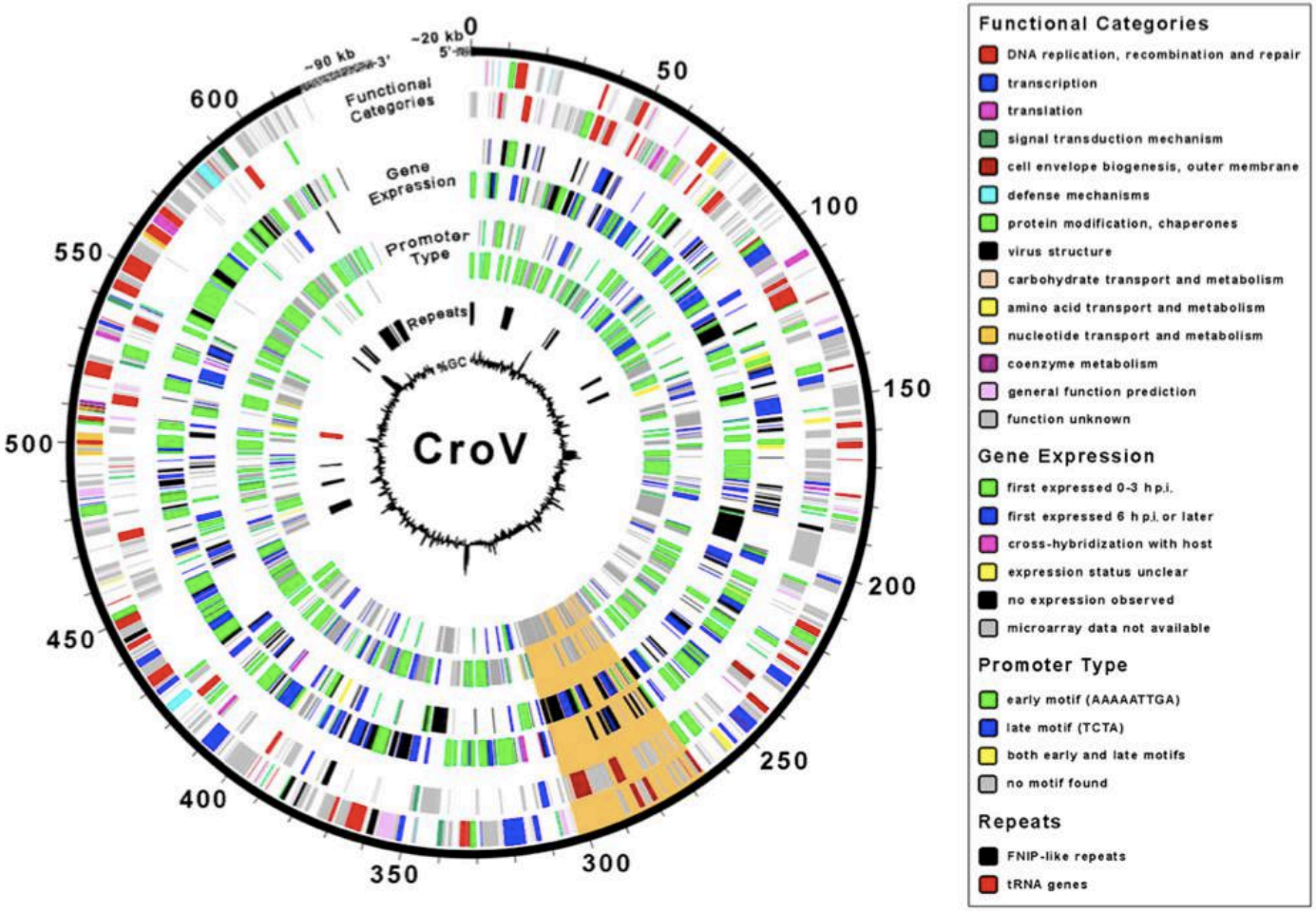


-<2 kb, circular ssDNA genome, encoding only for two proteins

- Viral life cycle:

1. Virus penetrates into the host cell.
2. Uncoating, and release of the viral genomic ssDNA into the nucleus.
3. The ssDNA is converted into dsDNA with the participation of cellular factors.
4. viral mRNAs are transcribed and translated to produce viral proteins.
5. Replication may be mediated by a "Rep-like" protein, and would occur by rolling circle
6. These newly synthesized ssDNA can either
   a) be converted to dsDNA and serve as a template for transcription/replication
   b) be encapsidated by capsid protein and form virions released from the cell by budding

Giant viruses - *Cafeteria roenbergensis*

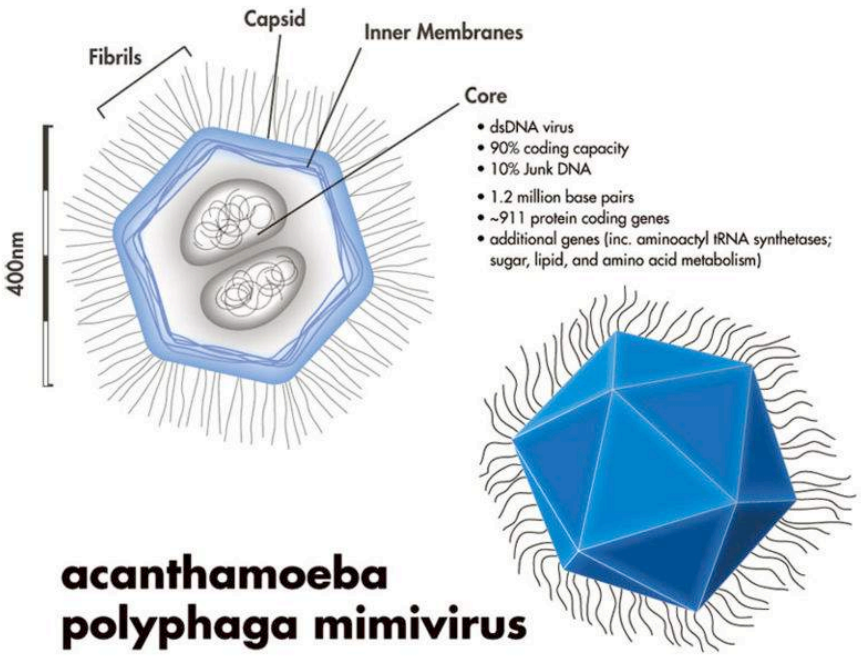


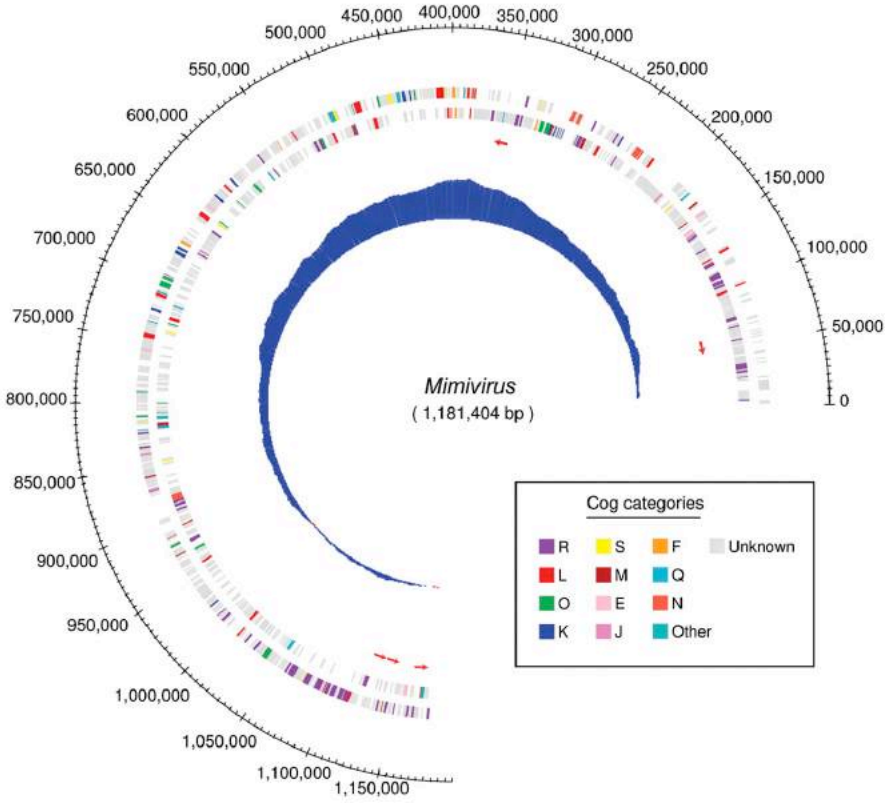
-~730 kb dsDNA genome, ~550 genes, some involved in translation, others in DNA repair

- 5% of the genome is repetitive DNA and a huge chunk of the genome is of bacterial origin

(Fischer et al. (2010) *PNAS*)

Giant viruses - *Acanthamoeba polyphaga mimivirus*



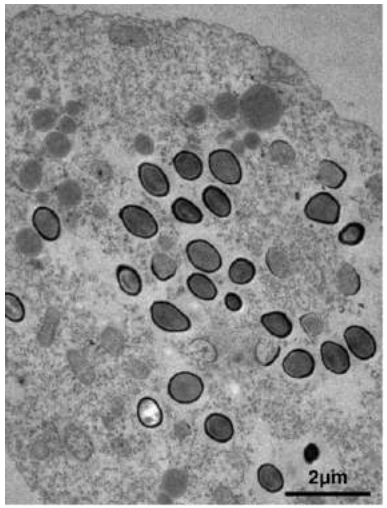


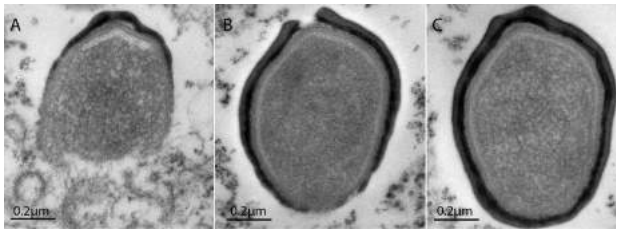
-~1.2 Mb dsDNA genome, 981 genes, many tRNA-synthases, genes involved in transcription, DNA repair

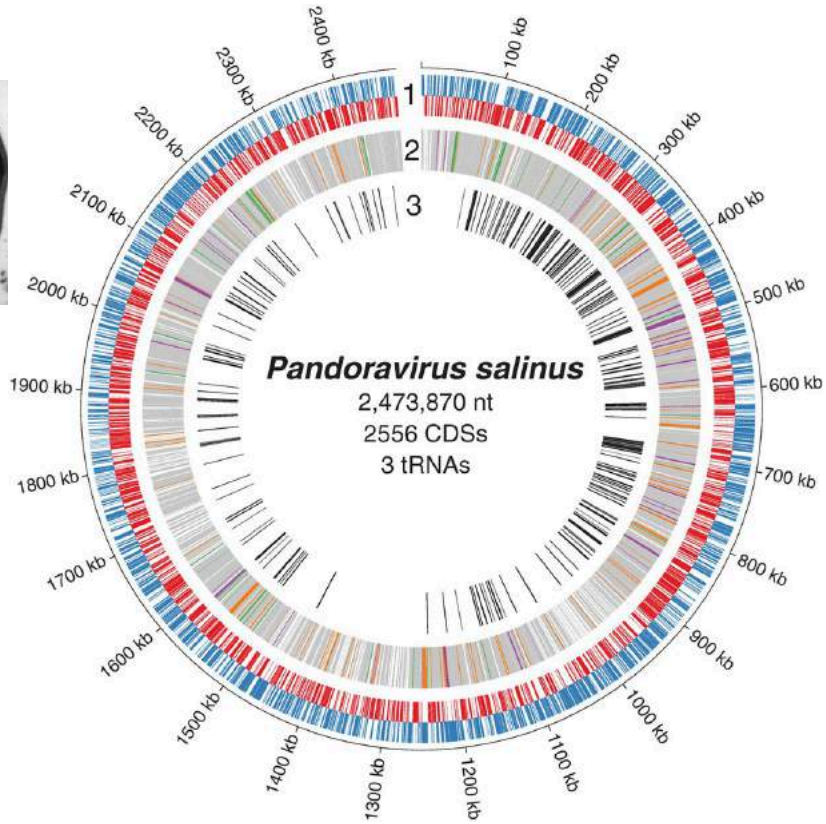
- the genomic boundary between viruses and real cellular organisms is opaque (or non existent): these viruses are larger than some prokaryotic viruses and encode a complex replication machinery

(Raoult et al. (2004) *Science*)

Giant viruses – *Pandoravirus salinus*



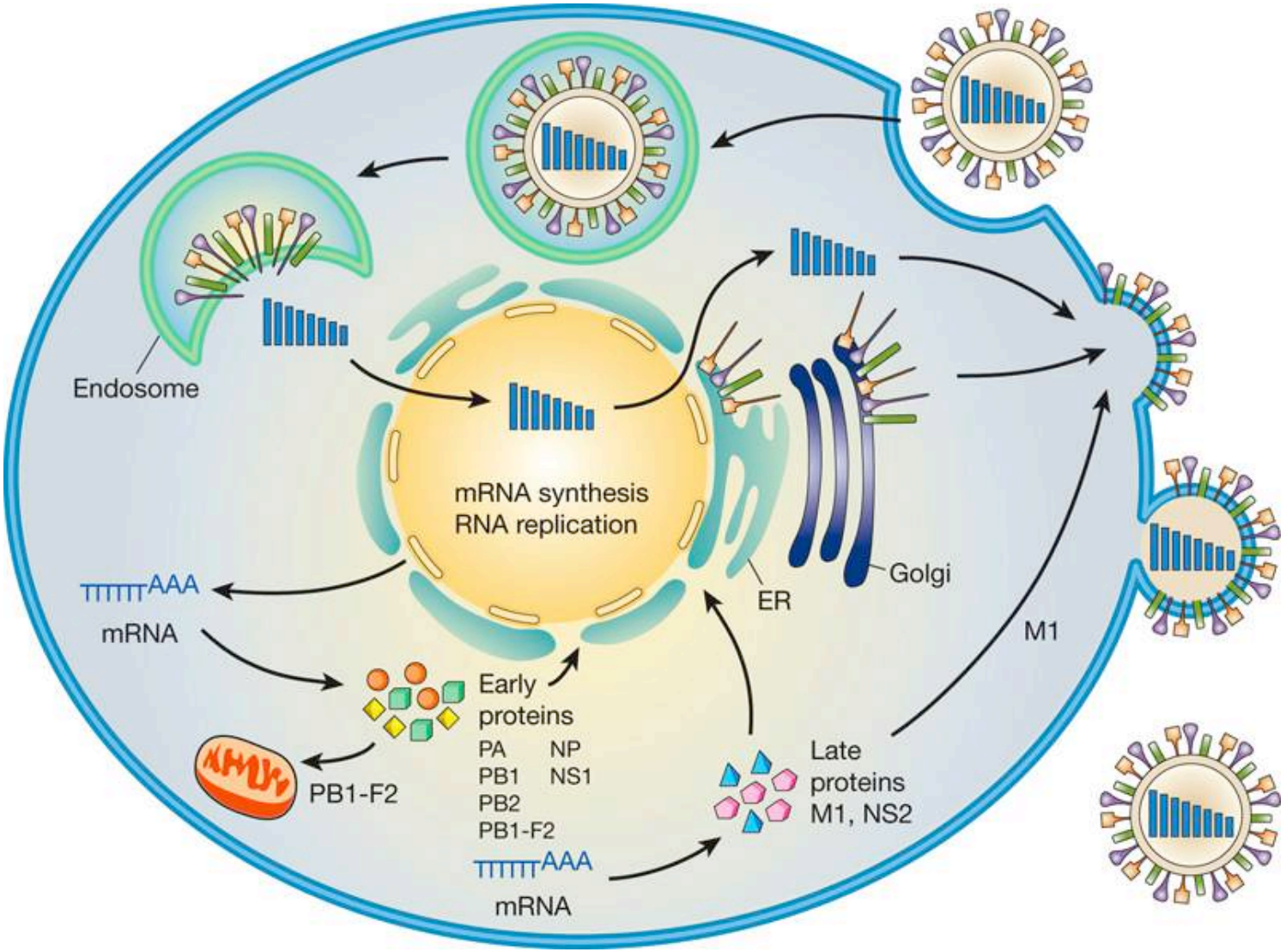




- the virus of *Acanthamoeba castellani*, discovered in Chile.
- forms 1um long, 0.5um wide particles
-~2.5-2.8 Mb dsDNA genome, 2556 hypothetical protein coding genes
- BUT: most of these (93%) have no homologs in other organisms (unusual even in viruses), thus it is possible that their translation is unusual

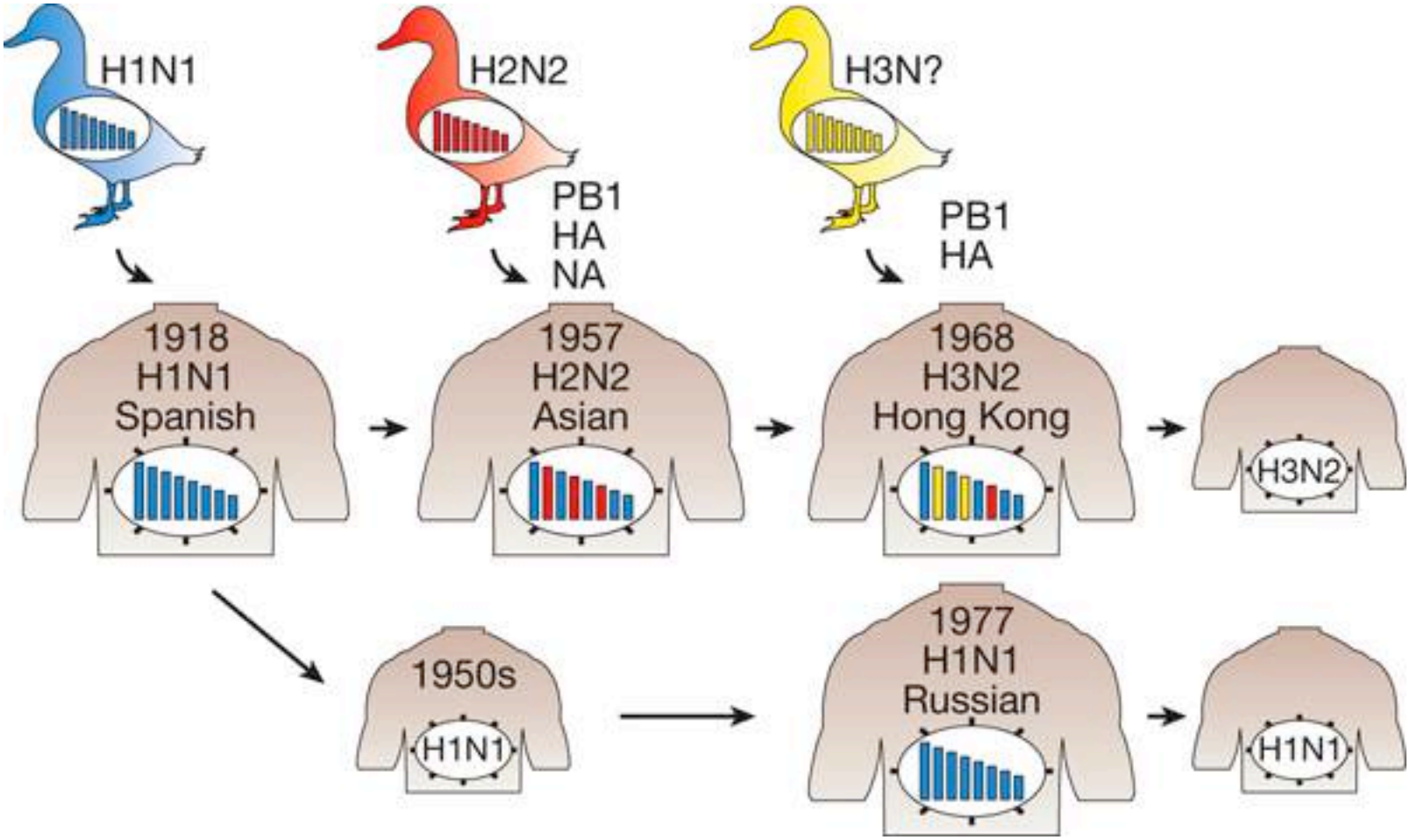
(Philippe et al. (2013) *Science*)

Life cycle of flu viruses



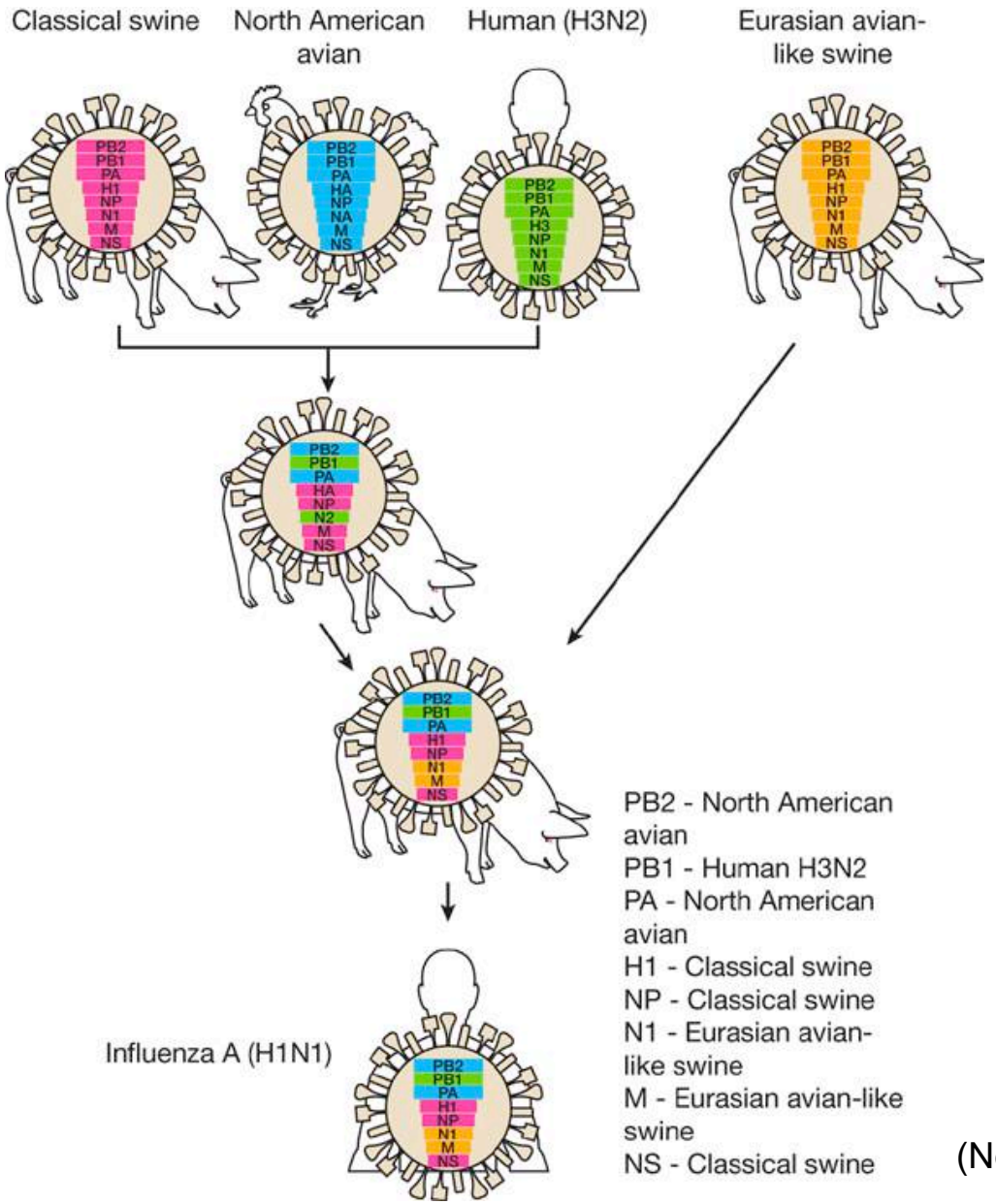
(Neumann et al. (2009) *Nature*)

Evolution of flu viruses through reassortation



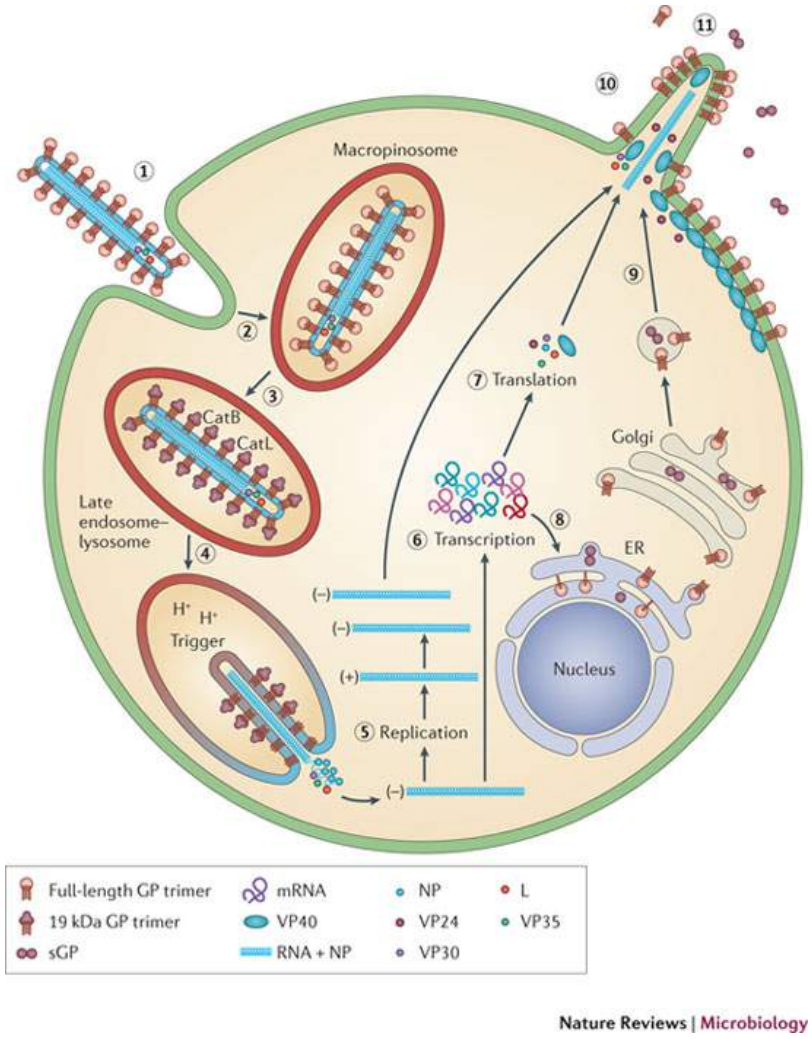
(Neumann et al. (2009) *Nature*)

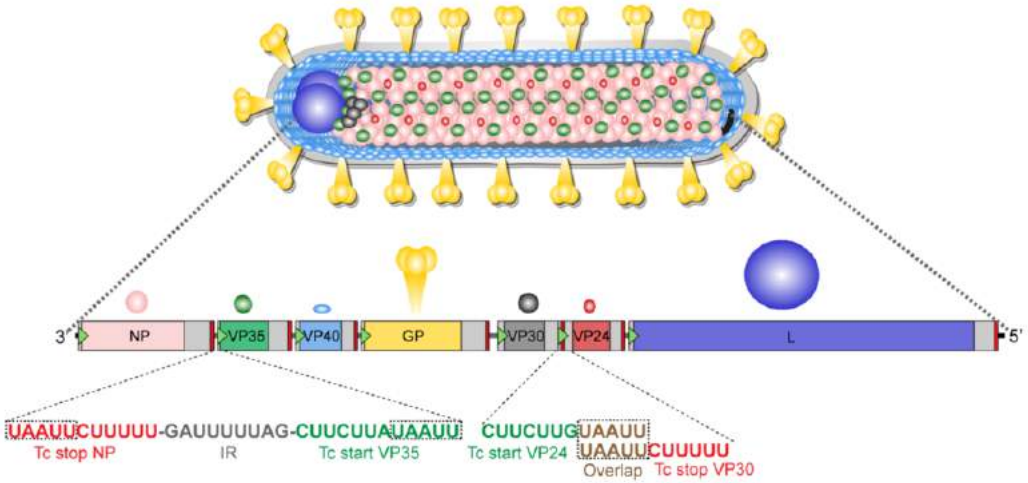
Influenza – the origin of the 2009 A(H1N1) strain



(Neumann et al. (2009) *Nature*)

The life cyle of the ebola virus

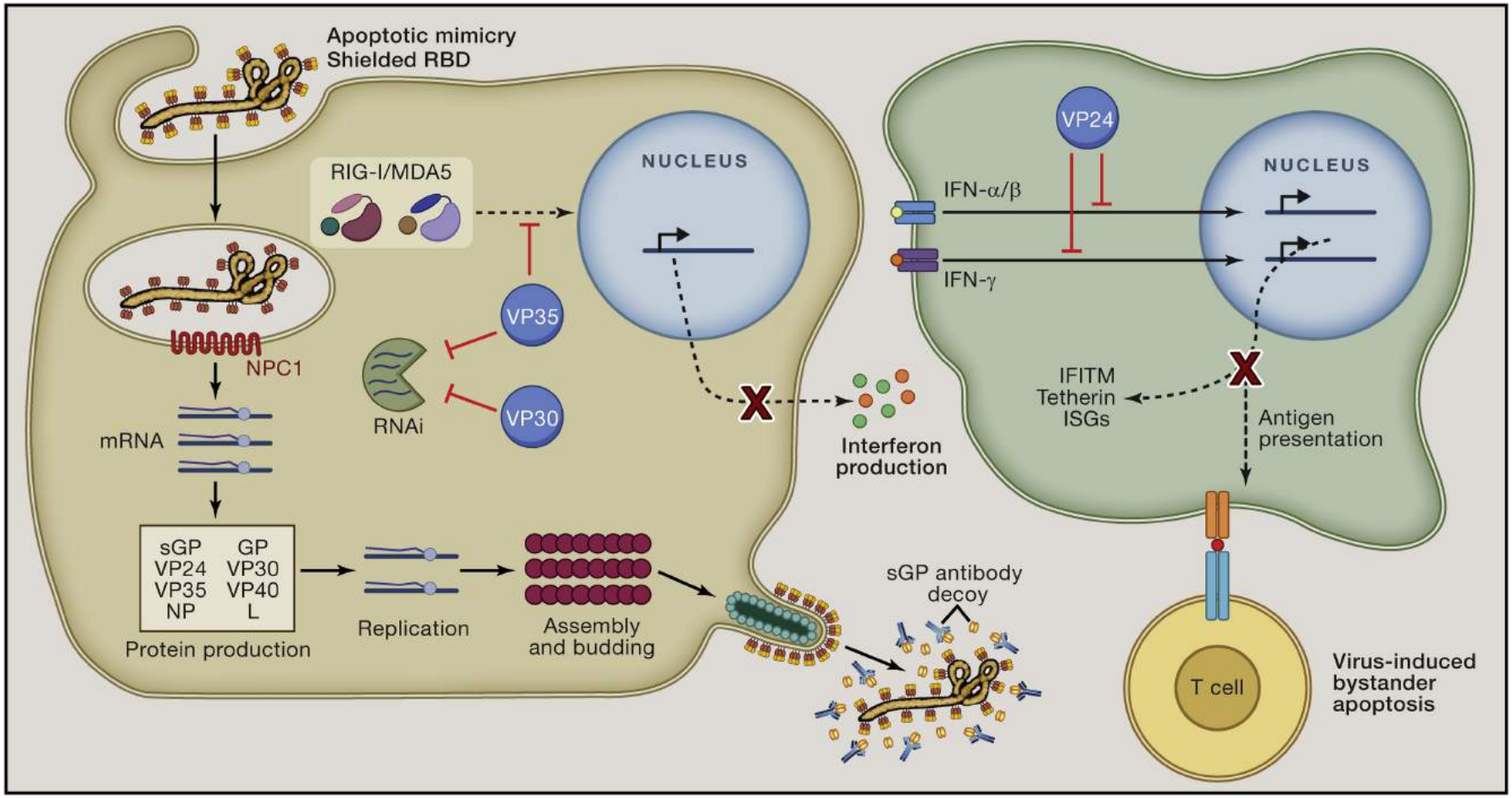




- (-) stranded RNA-genome
- Encodes 7 genes
- VP35, VP30, VP24 have a role in the suppression of the immune (IFN, RNAi)
- VP40 – matrix protein
- L – polymerase
- NP – nucleoprotein
- GP - glycoprotein

(White and Schonberg, 2012)

The life cyle of the ebola virus



(Misasi and Sullivan (2014), *Cell*)

The genome of the Zika virus